MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

Working Paper 284 December 1985

# Construction and Refinement of Justified Causal Models Through Variable-Level Explanation and Perception, and Experimenting

Richard J. Doyle

**Abstract:** The competence being investigated is *causal modelling*, whereby the behavior of a physical system is understood through the creation of an explanation or description of the underlying causal relations.

After developing a model of causality, I show how the causal modelling competence can arise from a combination of inductive and deductive inference employing knowledge of the general form of causal relations and of the kinds of causal mechanisms that exist in a domain.

The hypotheses generated by the causal modelling system range from purely empirical to more and more strongly *justified*. Hypotheses are justified by *explanations* derived from the domain theory and by *perceptions* which instantiate those explanations. Hypotheses never can be *proven* because the domain theory is neither complete nor consistent. Causal models which turn out to be inconsistent may be repairable by increasing the resolution of explanation and/or perception.

During the causal modelling process, many hypotheses may be partially justified and even the leading hypotheses may have only minimal justification. An experiment design capability is proposed whereby the next observation can be deliberately arranged to distinguish several hypotheses or to make particular hypotheses more justified. Experimenting is seen as the active gathering of greater justification for fewer and fewer hypotheses.

# Table of Contents

# Chapter 1

# Introduction

## 1.1 The Problem, the Motivation, and the Domains

One of the most important skills underlying common sense is the ability to recognize, describe, and reason about regularities and dependencies in the world in terms of *causal* relations. Causal descriptions enable us to generate useful explanations of events, to recognize the consequences of our actions, to reason about how to make things happen, and to constrain our theorizing when unexpected events occur, or expected events do not occur. The ability to identify causal relations and to construct causal descriptions underlies our ability to reason causally in understanding and affecting our environments.

Imagine waking up in the morning to find the refrigerator door ajar. It is not hard to understand why the food is spoiled, even if one is not quite awake. People commonly turn down the volume control on the home stereo before turning the power on, anticipating and knowing how to prevent a possibly unpleasant jolt. When the flashlight does not work, we will sooner or later check the batteries.

The causal reasoning tasks implicit in the scenarios above include *explaining* the spoiled food; *predicting* a possible consequence of flipping the power switch on the stereo and *planning* to prevent that event; and *diagnosing* the faulty behavior of the flashlight.

Before any of this useful causal reasoning can take place, the causal descriptions which support it have to be constructed or learned. Often they are not available, complete and appropriate for the problem at hand. Instead these causal descriptions have to be proposed on the basis of observations of the system to be understood and perhaps knowledge already in hand about such systems. The problem of causal modelling – the recognition and description of causal relations – in the context of a causal reasoning task is the *competence* which is the subject of this thesis investigation.

The class of domains which will test the principles emerging from this research effort, and the implementation based on them, is that of physical devices and gadgets – simple designed physical systems. Examples are cameras, showers, toasters, and air conditioners.

Causal reasoning is very important in domains like these. A photographer knows he or she has just wasted time setting up a shot when he or she notices that the lens cap was on when the shutter clicked. The light which would have created an image on the film never reached the film. Someone

drawing a bath knows he or she can leave the room to attend to other affairs because the safety drain will prevent overflow of water onto the floor. Someone who drops some bread into the toaster and starts reading the morning paper may, after a while, not having heard the toast pop up, check if the toaster is plugged in.

The reasoning outlined in each of these scenarios is supported by a causal model of the particular physical system involved. I propose to show how these causal models can be acquired in the context of a causal reasoning task. The causal reasoning task serves both to constrain the learning task, and to demonstrate that the learning does indeed produce useful causal descriptions.

## 1.2 The Issues

There are several issues associated with the causal modelling competence. This research will address these issues and attempt to provide principled solutions to them. Indeed, any performance results I present will be suspect if they are not accompanied by such general arguments.

- When/how to construct multi-level causal descriptions/explanations?

  Physical systems usually admit to description at more than one level of abstraction. Observations also come at various levels of granularity. I want my learning system to be able to construct causal descriptions at different levels of abstraction. I also want it to be able to determine when it might be useful/necessary to examine or explain a physical system at a finer level of detail. The causal reasoning task motivating the learning task can help to determine to what level explanation and perception should go.

- How to learn from experiments?

  The ability to design experiments to test and distinguish hypotheses is useful to a learning system. My example-driven inductive inference method can benefit from an analysis of the current set of hypotheses to determine what next example can maximally distinguish the hypotheses. Also, my domain theory can be exploited to determine what next observation can lead to more justified causal hypotheses.

- What kinds of causal relations exist in the physical system domain?

  This issue is a knowledge engineering issue. The result of the knowledge enginnering is the domain theory that guides the learning system. I assume that this domain theory is not necessarily complete or consistent. I have looked to the field of physics to ensure that the domain theory has some basis, and is not simply *ad hoc*.

- How to learn new compositions of causal relations for the domain?

  This issue concerns extending the domain theory which drives learning. I show how new, *composed* causal relation types for the domain can be acquired through an explanation-based learning technique. This approach does not support acquisition of new *primitive* causal relation types.

- How can empirical and analytical learning techniques complement each other?

  Although this research was originally motivated by the specific problem of learning causal models, some issues have arisen that are relevant to learning in general. In particular, my

learning system combines inductive and deductive inference and has both empirical and analytical learning components. I attempt to step back from the specific interactions of inference methods and learning paradigms in my system, and consider whether there are principled ways of combining these approaches.

- What is a good representation for causal relations?

  Choice/design of a representation language is an important step in the construction of any learning system. I have looked to the field of philosophy to come up with a model of causality; this model has guided my design of representations for causal relations. The representations, in turn, have more or less suggested inference methods, both inductive and deductive, which drive the learning of causal models.

## 1.3 The Learning Tasks

The primary learning task of my system is to construct causal models of physical systems to support a given causal reasoning task. The models are constructed from observations of the physical systems and from knowledge of the general form of causal relations and of the kinds of causal relations that exist specifically in the physical system domain.

The primary learning task is, more formally,

Given a causal reasoning task concerning a physical system, construct a causal model which supports the causal reasoning task, as follows:

- Given *structural* and sequence-of-events *behavioral* descriptions of a physical system,
- Hypothesize causal relations which can account for the observed behavior, and when possible,
- Justify proposed causal relations by determining if their types of interaction and supporting system structure indicate a known causal mechanism.

This primary learning task is supported by a background learning capability. The task of this second learning component is to identify and generalize compositions of the known causal mechanisms in the causal models produced by the primary learning component. These new, composed causal mechanisms can then be incorporated into the domain theory which drives the first learning component, enhancing its performance.

The secondary learning task is, more formally,

Given a causal model,

- Identify within the causal model new compositions of known causal mechanisms.
- Generalize these new composed causal mechanisms by generalizing their components without violating the constraints of the composition.

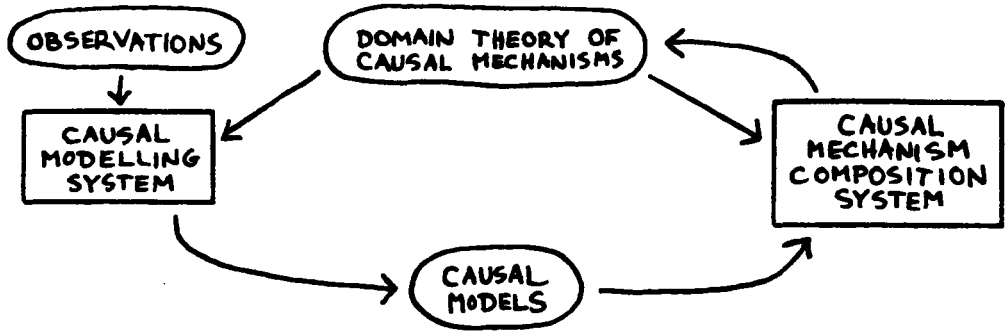


Figure 1.1: The Learning Systems

- Add the composed causal mechanisms and their generalizations to the set of known causal mechanisms.

This secondary learning system falls into the class known as *Learning Apprentice Systems*. These are learning systems which lurk in the background of other AI systems, extending their knowledge bases by monitoring and analyzing the activities of the systems (and perhaps the users) whose shoulders they are looking over. In my setup, the Learning Apprentice System identifies and generalizes justified compositions of causal mechanisms in the causal models constructed by the primary learning system.

## 1.4 A Scenario – The Camera Domain

This section contains a brief outline of the steps my learning system goes through in constructing a causal model of a camera to support the causal reasoning task of how to control the exposure of photographs (a planning problem). I give some hints of the kinds of knowledge it brings to bear and the inferences it makes in performing this learning task.

The learning system's initial observation of the camera includes a description of the various physical objects that make up the camera system and structural and geometrical relations between these objects.

*... The lens, the flash, and the release button are attached to the camera.*

*The aperture ring and the focus ring are attached to the lens.*

*The film is inside the camera. ...*

This *structural* description is complemented by a *behavioral* description which tells how the structural and geometrical relations, as well as the values of quantities of the objects, change over time.

*... Initially, the f-stop of the aperture ring is 5.6.*

*The distance of the focus ring is 3 feet.*

*The intensity of the flash is dark. ...*

*... Next, the intensity of the subject is bright.*

*The position of the release button is down.*

*The intensity of the flash is bright. ...*

*... Later, the film is outside the camera.*

*The film is overexposed. ...*

At first, almost any hypothesis about what could be affecting what is supported. The only constraints in force are that effects cannot precede their causes and that values and relations established by external actions must be primitive causes.

Already though, some hypothesized interactions have the earmarks of causal mechanisms that the learning system knows about. The simultaneous brightening of the flash and the subject could be light transmission. Covering up the flash might be a useful experiment.

Some of the proposed causal relations cannot be instantiated as instances of known causal mechanisms, either because the relevant mechanisms are unknown, or they can not be observed. For these uninstantiable causal relations, the learning system can still gather further empirical evidence by performing know-tweaking experiments. This evidence can strongly suggest, for example, that the setting of the aperture ring seems to affect the film exposure, while settings of the focus ring do not.

Eventually, when the learning system can distinguish or characterize its hypotheses no further, and there is strong empirical evidence for some causal relations, it should be motivated to look inside the camera. Perhaps mechanisms will be revealed which can justify the causal relations which have only an empirical basis. After opening the camera, the learning system can in fact instantiate a mechanical coupling between the aperture ring and the iris in the lens, indicated by the simultaneous changes in position or shape. It can also note a light path between the subject and the film and note that the iris and the shutter control the flow of light along this path.

As a final step, the learning system might note how the f-stop can affect the film exposure via interacting mechanical coupling and flow mechanisms. This complex mechanism might appear in other physical systems and should be incorporated into the learning system's domain theory.

This scenario reveals some of the sources of constraint my learning system employs and some of the reasoning it uses in constructing causal models. The number of hypotheses the learning system may entertain at various times is only suggested. A simplified version of the causal model of the camera which is the output of the learning system appears in the following figure.

The learning task cannot be considered complete until the constructed causal model can be shown to support the original planning problem of how to control the exposure of photographs. Planning involves finding sequences of operations which achieve a goal. When the causal model has revealed chains of causality which originate in available operations (e.g., setting the aperture ring) and terminate in the specified goal (values of the film exposure), then the learning task is complete. The causal model above can be used to generate plans for controlling exposure.

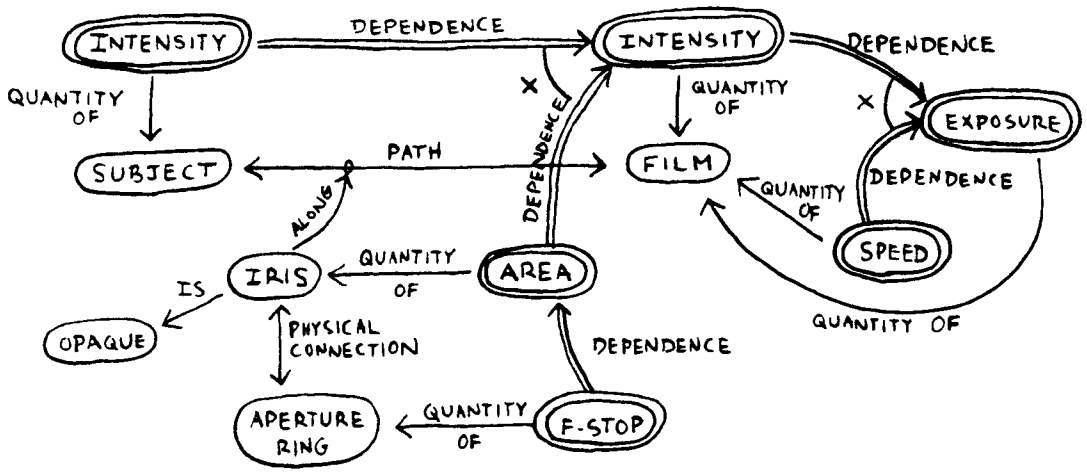


Figure 1.2: A Causal Model

# Chapter 2

# What is Causality?

In order to convincingly address the issues underlying this research into the causal modelling competence, I have looked to the field of philosophy for ideas on what exactly is the concept of causality. The theory of causality which I have adopted is due to Mackie [Mackie 74]. In his theory, causation has three aspects, *regularity*, *direction*, and *mechanism*. Regularity and direction are necessary explicit attributes of any causal relation, while mechanism is not. Regularity refers to the consistent repeatability of an observed co-occurrence of a cause (an event, a state, a relation, or some collection of them), and an effect (similarly defined). Direction refers to the cause being always somehow prior to the effect.

Causal relations which incorporate regularity and direction may be called the *associationist* causal relations. They are based purely in empirical evidence of an association that always has been observed to hold. However, some causal relations also incorporate a notion of *mechanism*, which gets at the question of why the effect should *necessarily* follow from the cause, or what is the tie, or process, by which the cause *produces* the effect. This notion of necessity is common to all causal relations, but the mechanistic causal relations are the only ones which make the justification for it explicit.

## 2.1 Causal Direction

Causal relations always have a direction. An effect never precedes a cause. This property of causal relations is based partly in *temporal* direction. If event $A$ occurs at time $t$, and event $B$ occurs at time $t + 1$, then $A$ must be the cause. But causal direction does not arise solely from temporal direction for often events which are *perceptually* simultaneous can be separated as to which is cause and which is effect.

The competence of identifying causal direction may also make use of knowledge (perhaps heuristic) about *primitive* causes. These are events which are taken to be always external inputs to a system. An example of an event which is a primitive cause is the action of an agent on a physical system. [1]

[1] This may not be a primitive cause in a domain where there are causal interactions among agents; e.g., one agent may "cause" another agent to perform some action.

Thus the direction of causality between two causally linked events can be determined in several ways:

1. The cause is the event which is temporally prior to the other.
2. The cause is the event known to be a primitive cause.
3. The cause is the event which is causally linked to an event which satisfies any of these three conditions.

This analysis in no way closes the book on the origins of causal direction. In fact, it somewhat begs the question because primitive causes are defined only extensionally, i.e., no way is given of identifying primitive causes not known *a priori*.

## 2.2 Mechanistic vs. Associationist Causal Descriptions

A causal description should have more than an empirical basis; it should do more than associate some set of conditions or changes – the cause, with another – the effect. A causal description also should provide a mechanistic explanation of why the observed interaction occurs. Such an explanation should answer questions like: "What is the medium which supports the causation?", "What kinds of barriers can prevent the interaction from taking place?", and "What class of causal mechanism or process does this interaction fall into?"

An associationist explanation of why the film in a camera came out blank might be: "One of the conjunctive preconditions for the dependence between the brightness of the subject and the intensity value on the negative, namely the lens cap being off, was unsatisfied." A mechanistic explanation might be: "The camera works by providing a controlled channel through which light may flow from the subject to the film. In this case, the lens cap acted as a barrier, preventing light from reaching the film."

The associationist explanation can be justified by noting that the film comes out blank/non-blank whenever the lens cap is on/off the lens. But there may be other statements whose truth-value is correlated, perhaps coincidentally, with the exposure or non-exposure of the film. Perhaps the camera was always on/off a tripod when the film came out blank/non-blank. The associationist description cannot further distinguish the lens cap and the tripod as being possible preconditions.

However, a mechanistic explanation of the lens cap in terms of a barrier to a light path does distinguish it *a priori* from the tripod. Such an explanation of the tripod is not forthcoming.

Associationist descriptions are restricted to aspects of the form of a causal relation: conjunctive or disjunctive, necessary or sufficient, the number of contributions, the sign of a dependence. They can never be supported by anything more than empirical evidence. On the other hand, mechanistic descriptions involve concepts like media supporting causation and barriers inhibiting causation. They provide hooks into domain knowledge, knowledge that can *justify* the proposal of a causal relation by revealing the *a priori* known mechanism that underlies the causation. An empirical, associationist causal description can never be justified in this way.

## 2.3 What is a Causal Mechanism?

I have spoken of mechanistic descriptions of causal relations without being entirely clear on what I mean by "mechanistic." Roughly, I mean causal descriptions which say something about what is the *tie* between the cause and the effect? What is it about the cause and perhaps the immediate environment which *produces* the effect?

These questions get at what philosophers call the *necessity* of causes. This necessity, if detectable, gives license to the inference that the cause, in the right circumstances, always results in the effect.

It is exactly this necessity which a mechanistic causal description is intended to reveal. This is the primary reason why a purely associationist representation of causal relations, which lacks this necessity, is inadequate. Such a representation may capture the form of a causal relation, but it says nothing about why the effect should and does follow from the cause.

A causal mechanism, to be called a mechanism at all, must describe some kind of *process* by which the cause produces the effect. Examples of processes are flow of material, transfer of momentum from one object to another, expansion of a material, transformation of electrical energy to light, heat, or motion. Processes, in turn, need some kind of *medium*, or structural link between the cause-object and the event-object, through which the interaction occurs. For example, flow of material requires a channel of some sort while transfer of momentum needs a physical connection between the objects. A complementary notion to medium is that of *barrier*. Whereas a medium is the structural link which enables a causal mechanism, a barrier is a disruption of the structural link which disables the causal mechanism. A channel can be nullified by a blocking of some kind which prevents flow. A physical connection which is broken cannot transfer momentum.

Processes reveal that there is considerable continuity between perceptually distinct causes and effects. They remove the mysteriousness that sometimes might be associated with a causal link. Processes typically describe a kind of persistence as well; some quality or form, present in the cause, is still present in the effect, although this persistence may not be always immediately perceivable.

A few examples of causation in physical systems and their mechanistic descriptions are illustrative here: A flashlight is turned on and a wall across the room brightens. The mechanistic explanation is that the flashlight emits light which flows through the intervening space to fall on the wall and brighten it. The cause-event and the effect-event are the changes in brightness of the flashlight and the wall, respectively. The medium is the straight-line, unobstructed path between them. The mechanism itself, or the process, is the continuous flow of light radiation from the flashlight to the wall. Finally, there is a persistence, or conservation of form, across the cause and effect; both involve changes in brightness.

Another example. The faucet in a shower stall is turned and momentarily, water emerges from the shower head. The mechanistic explanation runs thus: The faucet is connected to a pipe which runs to the shower head. Turning the faucet opens a valve in the pipe, allowing water to flow to the shower head and emerge there. There are two causal mechanisms involved. First, there is a mechanical coupling between the faucet and the valve. Turning the faucet results in movement of the valve. A physical connection, perhaps a rod, acts as medium here. The valve itself is a removable barrier to the flow of water to the shower head. This flow is the second causal mechanism. The pipe is the medium through which the flow occurs. In both of these causal mechanisms the cause resembles the effect. A movement of the faucet at one end of a mechanical connection is transferred to the other end, resulting in movement of the valve. A change in amount of water at one end of a pipe is propagated to the other end, where water emerges from the shower head.

The flashlight, straight-line path, and wall make up a system of light transmission. The faucet, rod, and valve comprise a mechanical coupling. The pipe and shower head are only part of a larger system of water transport, which includes a water source somewhere up the line.

There are appropriate barriers which could have disabled each of these causal mechanisms. The flashlight would not have brightened the wall if an opaque card was held in front of it. Turning the faucet would not have moved the valve if the rod was broken. And water would not reach the shower head if the pipe was clogged.

Each causal interaction consisting of a cause-object, medium, and effect object helps to locate the boundaries on the closed system inside which events can affect one another. This is roughly the spatial part of Pat Hayes' notion of causal enclosure, which he calls *history* [Hayes79]. What barriers do, essentially, is change the boundaries on the closed systems within which causal interactions occur. The flashlight does not stop emitting light, but it only brightens the card. Turning the faucet does move the rod, but only up to the break. And water flows to the clog but no further.

To summarize the notion of causal mechanism developed in this section: A causal mechanism describes how a cause-event propagates through some medium via some process, in the absence of relevant barriers, to produce an effect-event. The identified cause, medium, and process comprise a sufficient causal explanation for the observed effect, in that they provide a full accounting of why the cause, in these such and such circumstances, should result in the effect. This elaborate description goes further than any associationist causal description could; it reveals what is the tie, the philosophers' necessity, between cause and effect.

# Chapter 3

# Representations for Causality and Methods for Generating Empirical and Justified Causal Hypotheses

All learning systems are ultimately limited by the representation languages they use. A learning system cannot learn what it cannot represent. An important part of this research is the design of representations for causal relations.

I present both associative and mechanistic representations for causal relations, and inference methods based on them which perform the task of constructing causal models. The associative representations describe the general form of causal relations and support empirical, inductive inference methods. The mechanistic representation, on the other hand, supports the description of specific types of causal mechanisms which comprise a domain theory of causal relations for physical systems. The inference method which operates on this representation is deductive, and can generate justified hypotheses of causal relations, which the inductive methods can never do.

Although I argue that the mechanistic representation and its deductive inference method are preferable, nevertheless the associative representation and the inductive inference method have their place. This is because the domain theory driving deductive inference has inadequacies. For example, the domain theory is not considered to be complete. There may be causal mechanisms which operate in the domain which are unknown. The deductive method would fail utterly in its attempt to recognize an instance of an unknown causal mechanism. On the other hand, the inductive method, which does not utilize the domain theory, is unaffected by its incompleteness, and could still generate hypotheses purely from empirical evidence.

## 3.1 A Boolean Representation for Causality

I begin my survey of representations for causality and inference methods which operate on them by considering the work of John Stuart Mill, a philosopher who worked on the causal induction problem [Mackie 67]. In his scheme, the simplest representation for a causal relation is as in Figure 3.1.


Figure 3.1: A Simple Boolean Representation for Causality

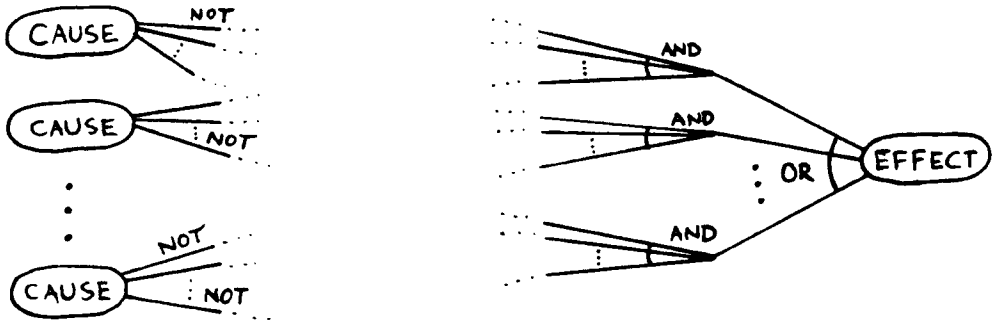


Figure 3.2: A Better Boolean Representation for Causality

A single cause (a condition, an event) results in an effect.

In an inductive context, the only constraints available for exploitation are regularity and direction. If the effect does not occur every time the suspected cause does (but not necessarily vice versa), then the correct cause has not been found.

This representation is too simplistic for several reasons. Causes typically do not cause in a vacuum. There is usually a set of relevant enabling preconditions which must be satisfied to bring about an effect. Or, there may be several disjoint ways in which an effect can be produced. Finally, it may be the absence, rather than the presence of some condition which brings about an effect.

A more sophisticated representation of causal relations would allow arbitrary conjunctions, disjunctions, and negations in the cause. Also, the cause may be considered necessary, or sufficient, or both for the effect.

### 3.1.1 Mill's Methods of Causal Induction

Mill formalized a set of methods of causal induction which operate on this general boolean representation for causal relations. Perhaps the most interesting aspect of his work is his careful analysis of the interactions between assumptions, observations, and inductive inferences. A more accessible account of Mill's methods may be found in the appendix to Mackie's book on causation [Mackie 74]. The examples to come are drawn from Mackie.

To illustrate Mills's methods of causal induction, consider the following example. Suppose that $A$, $B$, $C$, and $D$ are possible causes [1] of $E$ and we have the following observation.

[1] There is an unaddressed issue lurking here – namely the identification of possible causes. How are $A$, $B$, $C$, and

| OBSERVATION | $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|---|
| Example 1 | T | T | F | ? | T |
| Example 2 | F | T | ? | T | F |

There is one positive example in which the effect $E$ occurs, and one negative example in which it does not. The inductive task is to determine what cause for $E$ was present in the positive example and absent in the negative one.

If we assume that some single, unnegated condition is necessary and sufficient as a cause for $E$ then we can conclude that $A$ is this cause. $B$ and $D$ are eliminated because their presence in the negative example, when the effect did not occur, shows they are not necessary. $C$ is eliminated because its absence in the positive example shows it is not sufficient.

If instead we assume that a cause may be negated, i.e., the absence of some condition may be the relevant cause, then the observation above cannot eliminated $C$ and $D$. We need a stronger observation which shows that $C$ was false in both examples and $D$ was true in both.

If we relax our assumption about the form of the cause to allow conjunctions (with unnegated conjuncts), then the hypotheses $AB$, $AD$, and $ABD$ must be kept under consideration. We can conclude that $A$ is necessary, but perhaps not sufficient.

The interactions between assumptions, observations, and inductive inferences become more complex as we admit various combinations of negations, conjunctions, and disjunctions. In general, the weaker the assumption (the greater the number of possible forms for the cause), the more underconstrained the conclusion, given the same observation.

The manipulation of assumptions and inferences is familiar to the AI community as non-monotonic reasoning [Doyle 79]. The role of assumptions in inductive reasoning specifically also has been treated [Utgoff 85]. What is impressive about Mill's work, and perhaps sobering for AI researchers, is that he was attentive to the need for making assumptions explicit and was able to analyze the consequences of manipulating them. In the context of his methods of causal induction, he made a quite complete treatment of these issues.

## 3.2 A Representation for Causality Based on Quantities and Functional Dependencies

In the representation scheme for causal relations outlined above, the conditions, events, and/or changes which make up causes and effects are represented by statements which are true or false. This is an inadequate representation for describing the physical systems which my learning system will investigate. In particular, there is no easy way to describe the continuous properties of objects in such systems, properties such as height, temperature, velocity, etc. In the boolean representation, we would have to say that at any time, one possible value of a quantity was true and all others were false. An awkward representation.

Instead, I have adopted Forbus' qualitative representations for quantities, their value spaces, and functional dependencies between them [Forbus 84]. These form the basis for another associative representation for causality which supports inductive inference. This representation is designed to easily capture the kinds of continuous phenomena that occur in physical systems, which the boolean representation could not.

$D$ brought under suspicion as possibly participating in a cause for $E$ in the first place? The inductive methods given in this section provide no handle on this issue. I return to it later.

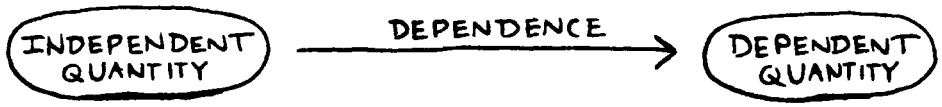


Figure 3.3: A Simple Representation for Causality Based on Functional Dependencies

### 3.2.1 Another Set of Methods of Causal Induction

The task now is to develop a set of inductive methods which operate on the representation for quantities and dependencies. The inductive method presented in this section is an expanded version of work done in my master's thesis [Doyle 84].

In this new representation, the simplest kind of causal relation is a single direct proportionality relation between two quantities: $y = p(x)$. To make life as simple as possible, we can assume further that the proportionality relation is *monotonic* and that the zero values of the two quantities coincide (i.e., $y = 0$ when $x = 0$).

The single direct proportionality assumption under the functional dependencies representation of causality is analogous to the single unnegated cause assumption under the boolean representation. The inductive task is also correspondingly simple.

For example, consider the following set of observations. Observations under the new representation consist of the signs of quantities. [2]

| OBSERVATION | *A* | *B* | *C* | *E* |
|---|---|---|---|---|
| Example 1 | + | − | 0 | 0 |
| Example 2 | + | − | + | + |
| Example 3 | + | 0 | + | + |

Under the single direct proportionality assumption, causes (independent quantities) are those quantities whose signs are always the same as the sign of the effect (dependent quantity). In the example above, with this assumption, only *C* could be a cause of *E*.

This representation of functional dependencies is too simplistic. Just as we were forced to consider various compositions of negation, conjunction, and disjunction under the boolean representation, now we must consider compositions of functional dependencies under ***negation***, ***multiplication***, and ***addition***.

Allowing negation admits the possibility of *inverse* proportionality relations: $y = -p(x)$. The following table shows how the signs of quantities are transformed under negation.

| *sign* | $-sign$ |
|---|---|
| 0 | 0 |
| + | − |
| − | + |

Thus a negative dependent quantity can now result in two ways. A direct dependence and a negative independent quantity, or an inverse dependence and a positive independent quantity. This

[2] Actually, the qualitative values of quantities, from which signs can be derived.

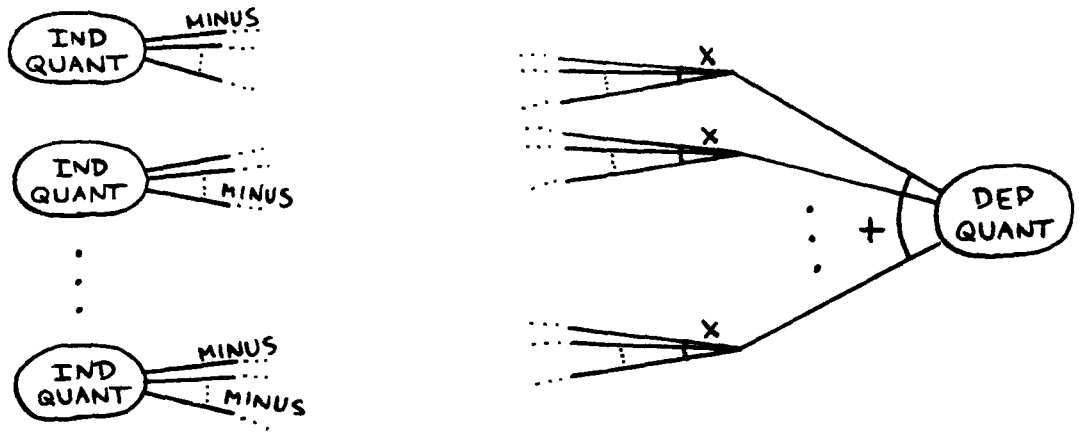


Figure 3.4: A Better Representation for Causality Based on Functional Dependencies

loosening of the assumptions about the form of the cause does not admit any further hypotheses for the observation above, but it might have. In particular, a quantity whose signs in the three observations above were $0,-,-$, respectively, would now have to be considered as a cause for $E$.

As long as we assume that only one independent quantity is affecting any dependent quantity at any time, the inductive task is still highly constrained. This is the analog of the single cause assumption for the boolean representation of causes and effects.

Once we relax the single-cause assumption by admitting multiplication and addition, things become more complicated.

We need to describe how contributions of several functional dependencies combining multiplicatively can affect the values of dependent quantities. The arithmetic form of this type of multiple functional dependence is $y = p_1(x_1) \times p_2(x_2) \times \cdots p_n(x_n)$. Combinations of the signs of two contributions under multiplication are listed in the following table.

| $sign_1$ | $sign_2$ | $sign_1 \times sign_2$ |
|---|---|---|
| 0 | 0 | 0 |
| + | 0 | 0 |
| − | 0 | 0 |
| 0 | + | 0 |
| + | + | + |
| − | + | − |
| 0 | − | 0 |
| + | − | − |
| − | − | + |

These two-operand combinations can be applied recursively [3] to determine how *several* contributions combine multiplicatively. Alternatively, this knowledge can be aggregated to the general case of $n$ contributions *a priori*, as has been done in the following table.

| $signs = 0$ | $signs = +$ | $signs = -$ | $sign_1 \times sign_2 \times \cdots \times sign_n$ |
|---|---|---|---|
| $\geq 1$ | $\geq 0$ | $\geq 0$ | 0 |
| 0 | $\geq 0$ | even | + |
| 0 | $\geq 0$ | odd | − |

[3] In any order, since multiplication is commutative and associative.

Using this knowledge of how contributions combine under multiplication we find that there are now more causal hypotheses that are consistent with the given observations, namely $A \times C$ and $(-A) \times (-C)$.

Now we consider multiple functional dependencies of the form $y = p_1(x_1) + p_2(x_2) + \cdots + p_n(x_n)$. Combinations of the signs of two contributions under addition are listed in the following table.

| $sign_1$ | $sign_2$ | $sign_1 + sign_2$ |
|---|---|---|
| 0 | 0 | 0 |
| + | 0 | + |
| − | 0 | − |
| 0 | + | + |
| + | + | + |
| − | + | 0 + − |
| 0 | − | − |
| + | − | 0 + − |
| − | − | − |

Again, we can compute combinations of $n$ contributions under addition by applying the two operand combinations recursively. Or we can aggregate the knowledge *a priori* as in the following.

| $signs = 0$ | $signs = +$ | $signs = -$ | $sign_1 + sign_2 + \cdots + sign_n$ |
|---|---|---|---|
| $\geq 0$ | 0 | 0 | 0 |
| $\geq 0$ | $\geq 1$ | 0 | + |
| $\geq 0$ | 0 | $\geq 1$ | − |
| $\geq 0$ | $\geq 1$ | $\geq 1$ | 0 + − |

Notice the ambiguity when contributions of opposite sign are added together. This indeterminism makes the additive induction problem inherently less constrained than that of negation or multiplication.

Armed with this knowledge of how contributions combine additively, we must now consider the following causal hypotheses, given the above observations: $A + B$, $A + B + C$, and $A + B + (-C)$.

Finally, we can consider the most general case, that of arbitrary compositions of negation, multiplication, and addition. Again, we can determine which compositions are consistent with the given observations by applying the appropriate combinations from the tables for negation, multiplication, and addition recursively. The hypotheses which now have to be admitted are $(A + B) \times C$, $(A + B) \times (-C)$, and $((-A) + (-B)) \times (-C)$.

### Transitions

This new set of inductive methods operate, like Mill's methods, through elimination of hypotheses. Hypotheses which do not satisfy constraints which specify the valid combinations of functional dependencies under negagtion, multiplication, and addition are removed from consideration. The constraints I have described so far concern how the *signs* of quantities *from individual observations* must combine under these operations. This is not the only source of constraint which can drive these inductive methods. Also of relevance are the ways in which the values of quantities *change* from one observation to the next. For example, a quantity whose value was positive in two successive observations may have actually decreased, stayed the same, or increased. These *transitions* also combine in well-defined ways under negation, multiplication, and addition. The value of specifying

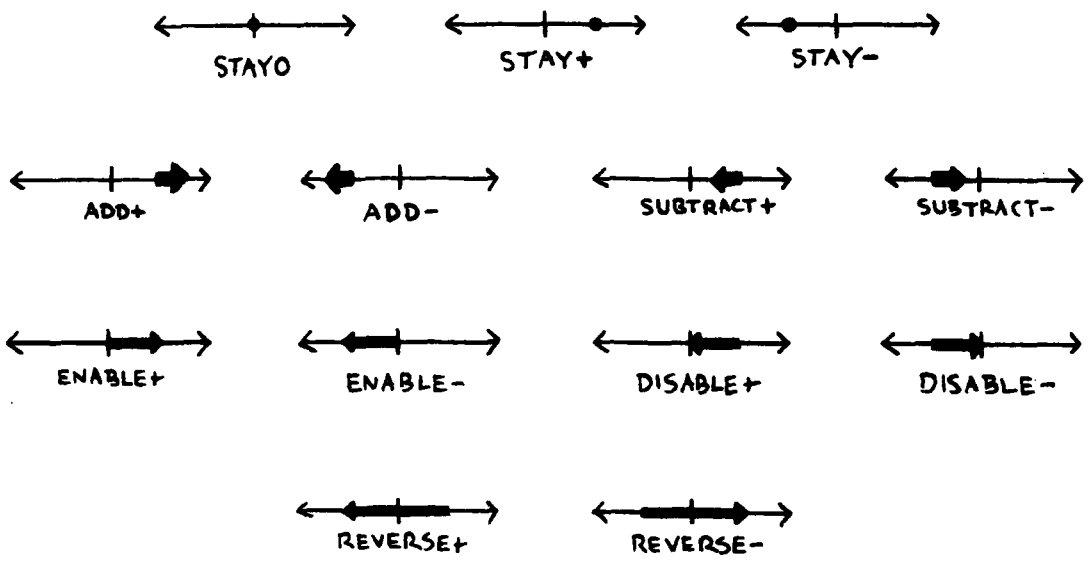


Figure 3.5: The Transitions

the exact transitions is that not all of them may be consistent with the given observations. Knowledge of transitions also can drive these eliminative, inductive methods.

The following table and above figure define the possible transitions of quantities between successive observations.

| Transition | Old sign | New sign | Direction of change | Abbreviation |
|---|---|---|---|---|
| STAY0 | 0 | 0 | 0 | ST0 |
| STAY+ | + | + | 0 | ST+ |
| STAY- | − | − | 0 | ST- |
| ADD+ | + | + | + | AD+ |
| ADD- | − | − | − | AD- |
| SUBTRACT+ | + | + | − | SB+ |
| SUBTRACT- | − | − | + | SB- |
| ENABLE+ | 0 | + | + | EN+ |
| ENABLE- | 0 | − | − | EN- |
| DISABLE+ | + | 0 | − | DB+ |
| DISABLE- | − | 0 | + | DB- |
| REVERSE+ | + | − | − | RV+ |
| REVERSE- | − | + | + | RV- |

The calculi which describe how these transitions combine under negation, multiplication, and addition appear in the appendix.

This knowledge of transitions and how they combine under negation, multiplication, and addition now can be used to make further inductive inferences. Observed transitions can be combined recursively according to the operations specified in the various causal hypotheses. The closures of the transitions of the causes (independent quantities) under the compositions specified in an hy-

pothesis must contain the observed transition of the effect (dependent quantity) for all observations; otherwise that hypothesis cannot be considered viable.

### 3.2.2 Examples of Inductive Inference

Consider again the earlier set of observations, now extended to include not only the signs of quantities in each observation, but also the transitions of quantities between observations.

| OBSERVATION | $A$ | $B$ | $C$ | $E$ |
|---|---|---|---|---|
| Example 1 | + | − | 0 | 0 |
| Transition 1→2 | AD+ | ST- | EN+ | EN+ |
| Example 2 | + | − | + | + |
| Transition 2→3 | AD+ | DB- | SB+ | SB+ |
| Example 3 | + | 0 | + | + |

Recall that one of the hypotheses that survived the earlier inductive inferences was $(A+B)$. Now we show that this hypothesis can be eliminated because the observed transitions are inconsistent with the form of (the composition of operators specified by) this hypothesis.

Quantity $A$ underwent an AD+ transition from example 1 to 2.

Quantity $B$ underwent a ST- transtion from example 1 to 2.

The combination of AD+ and ST- under addition is (AD+ SB- EN+ DB- RV-).

The transition of the dependent quantity $E$ from example 1 to 2 is EN+ which is in this closure. The hypothesis is still viable.

Quantity $A$ underwent an AD+ transition from example 2 to 3.

Quantity $B$ underwent a DB- transtion from example 2 to 3.

The combination of AD+ and DB- under addition is (AD+ SB- EN+ DB- RV-).

The transition of the dependent quantity $E$ from example 2 to 3 is SB+ which is not in this second closure. The hypothesis can now be eliminated.

Interestingly, the hypothesis $(A + B + C)$ does survive this analysis.

Recall that the set of possible transitions of $(A + B)$ from example 2 to 3 is (AD+ SB- EN+ DB- RV-).

Quantity $C$ underwent a SB+ transition from example 2 to 3.

The combination of (AD+ SB- EN+ DB- RV-) and SB+ under addition is (ST0 ST+ ST- AD+ AD- SB+ SB- EN+ EN- DB+ DB- RV+ RV-) – all possible transitions. This hypothesis cannot be eliminated.

Continuing this analysis, the final set of hypotheses consistent with the given observations is $C$, $A \times C$, $(-A) \times (-C)$, $(A + B + C)$, $(A + B) \times C$, $(A + B) \times (-C)$, and $((-A) + (-B)) \times (-C)$.

The knowledge about how signs and transitions combine under negation, multiplication, and addition makes up a kind of qualitative calculus which can be exploited to make inductive inferences.

This knowledge reveals two sources of constraint. There are a limited number of ways in which the signs of several contributions due to functional dependencies on a quantity can combine to produce the sign of that quantity. And there are a limited number of ways in which changes in the underlying contributions on a quantity can combine to produce changes in that quantity.

### 3.2.3 Qualitative vs. Quantitative Values

My set of inductive methods is driven by observations of the qualitative values of quantities. It is important to point out that qualitative values inherently carry less constraint than quantitative values. For example, an hypothesis involving two additive contributions, one of $+4$ and one of $-2$, say, can be eliminated as a possible explanation for a value of $+3$. On the other hand, if we know only the signs of these values, this hypothesis cannot be eliminated.

One valid characterization of my inductive method is that it is a qualitative version of a program like BACON [Langley et al 83], which performs quantitative function induction. Given quantitative observations, it is true that a quantitative induction method would outperform my qualitative method because it would make use of additional constraint to eliminate hypotheses more effectively.

However, I am assuming that my learning system will have access to qualitative values only. Specifically, the learning sytem will know the signs and the relative magnitudes of different symbolic values for quantities, i.e., whether the value of a quantity is greater than, less than, the same as, an earlier value of the same quantity or a value of another quantity of the same type. I think this is a realistic assumption for any AI system in a simulated (or real) visual environment. I do not assume the existence of perceptual equipment that can quantize observations.

### 3.2.4 Assumptions and Inductive Inferences Again

Mill's inductive methods can make inductive inferences only relative to assumptions about the form of boolean causal relations – in particular, whether negations, conjunctions, and/or disjunctions were to be admitted. Similarly, the set of inductive methods outlined above – based on a quantities and functional dependencies representation of causal relations – can make inductive inferences only relative to assumptions about how these causal relations might combine. Now the possible combination operators are negation, multiplication, and/or addition.

There are other assumptions we might make within the quantities and functional dependencies representation which can eliminate some hypotheses and hence constrain the induction problem.

For example, we might assume that there are never contributions in opposing directions (both $+$ and $-$). Or we might assume that in the case of additive contributions, the equilibrium state is stable and robust, i.e., that opposing contributions always relax to the equilibrium state. Another possibly reasonable assumption is that the equilibrium state is unattainable. Yet another is that multiplicands never can be negative, so that quantities can only be amplified and reduced via multiplication by non-negative values.

These possible assumptions can apply equally, in most cases, to the signs of contributions and their combinations and to transitions of contributions and their combinations. For example, the no equilibrium assumption applied to transitions of contributions means that transitions cannot cancel out under addition, i.e., combinations such as ADD+ + SUBTRACT+ $\rightarrow$ STAY+ are disallowed.

Another assumption, more subtle than the others, has to do with knowing the zero points of proportionality relations. As mentioned earlier, we assume by default that for any proportionality relation $y = p(x)$ (or $y = -p(x)$), $y = 0$ when $x = 0$. This need not be the case. Yet not knowing where these zero points are means it is impossible to determine whether a contribution is null, positive, or negative. [4]

Similarly, it is impossible to determine which of several transitions of the same direction of change is the correct one. Without some assumption about where zero points are, there is no alternative but to consider all possibilities for the signs and transitions of contributions – a severely underconstrained induction problem.

Yet another insidious assumption, one which is inescapable because of the impossibility of *confirming* hypotheses through induction, is that all relevant causes (independent quantities) are being considered. If this assumption is violated, there may be *no* hypotheses which are consistent with observations. This would be the case if the quantity $C$ was ignored in the above observations.

One reason why a relevant independent quantity might be ignored is because it cannot be observed at the given level of perceptual granularity. It is possible to propose hypotheses which include "ghost" or hidden quantities and test them with the inductive methods I have outlined. This, too, leads to a terribly underconstrained induction problem because, clearly, all possibilities have to be considered for the signs and transitions of hidden quantities.

All of these assumptions are posable and retractable in the non-monotonic reasoning sense. They constrain the number of hypotheses which have to be considered either by affecting the number of valid combinations of functional dependencies, or by affecting the number of valid interpretations of observations.

Still other assumptions are more or less hard-wired and are not amenable to non-monotonic treatment. One of these assumptions is that all proportionality relations are *monotonic*. Once the sign of one of these primitive functional dependencies is determined, it is known for all values of the independent quantity. Without this assumption, the best that could be hoped for with qualitative values is to try and isolate intervals in the range of values of the independent quantity within which the sign of the dependence does not change. A recent paper has shown how a conceptual clustering algorithm might be used for this task [Falkenheimer 85].

Another hard-wired assumption is one I call *omniresponsivity*. This assumption states that all perceivable changes in independent quantities result in perceivable changes in dependent quantities. Combined with the monotonicity assumption, this becomes *strict monotonicity* combined with a statement about perceptual resolution.

Making assumptions explicit is an important task in building any AI system. Inferences are always affected by assumptions and it is virtually impossible to design representations and procedures which operate on them which are assumption-free. Non-monotonic reasoning allows some assumptions to be manipulated. But even for those assumptions which are hard-wired, exposing them helps to reveal, in turn, limitations of the AI system that uses them.

## 3.3 A Representation for Causal Mechanism

In this section, I present a representation designed to capture the notion of mechanism which is central to the concept of causality in physical systems. Knowledge of the kinds of causal mechanisms

[4] This inference also depends on the *monotonicity* assumption for proportionality relations.

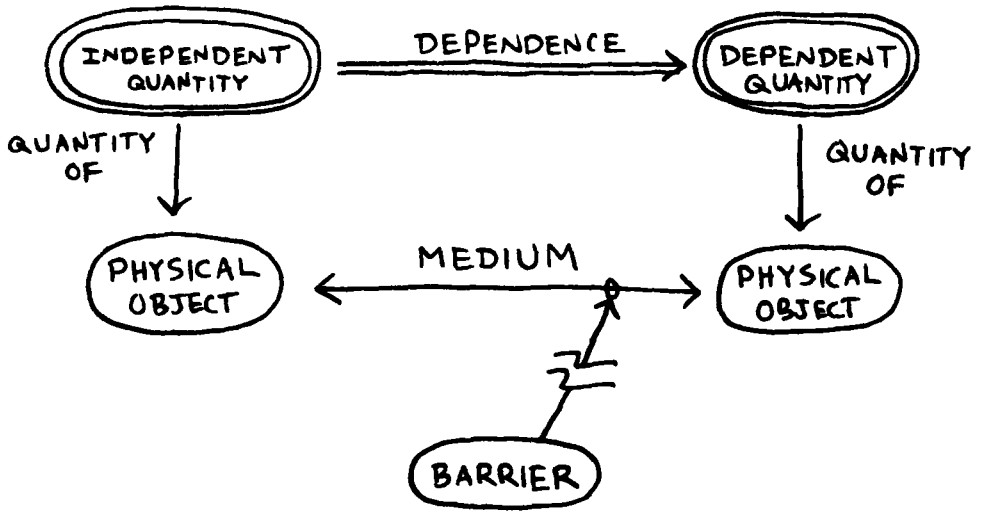


Figure 3.6: A Mechanistic Representation for Causality

that exist in the physical system domain can be described in this representation. The representation supports an alternative, more knowledge-intensive approach to the causal modelling problem involving a *deductive* method. This inference method generates *justified* hypotheses of causal relations as instances of known causal mechanisms.

The representation I have developed combines, at the level of form, the boolean representation and the quantities and functional dependencies representation for causal relations. In addition, and most importantly, it directly incorporates a mechanistic vocabulary. The set of ontological descriptors is extended to include terms like medium and process, instead of just domain-independent terms like precondition and dependence. Also, physical objects and quantities can be classified, which aids in the recognition of particular causal mechanisms.

This representation can be thought of as a template for causal mechanism schemata. The basic causal relation remains the functional dependence between quantities. Quantities are associated with physical objects, and all can be typed. Processes are essentially typed functional dependencies. There are two kinds of relevant preconditions. One is an enabling condition: there must be a structural link, a medium, between the cause object and the effect object. The other is a disabling condition: there must be no barriers which would disrupt the medium and decouple the cause and effect objects.

A mechanical coupling is an example of a causal mechanism relevant to the physical system domain which can be described within this representation.

A mechanical coupling is characterized by co-occuring motions of two objects. The relevant quantities are those that describe position. The medium between the objects is some kind of physical connection. A possible barrier is a break in the connection.

The representation for causal mechanism is an abstract structure with objects whose types are quantities, media, barriers, etc. Each causal mechanism described within this representation has the same structure but objects are specialized to particular classes of quantities, media, etc. The domain theory I present later includes not only a set of causal mechanisms for the physical system domain but also a generalization hierarchy for those mechanisms. Part of the domain knowledge is of the generalizations that are possible within the domain.

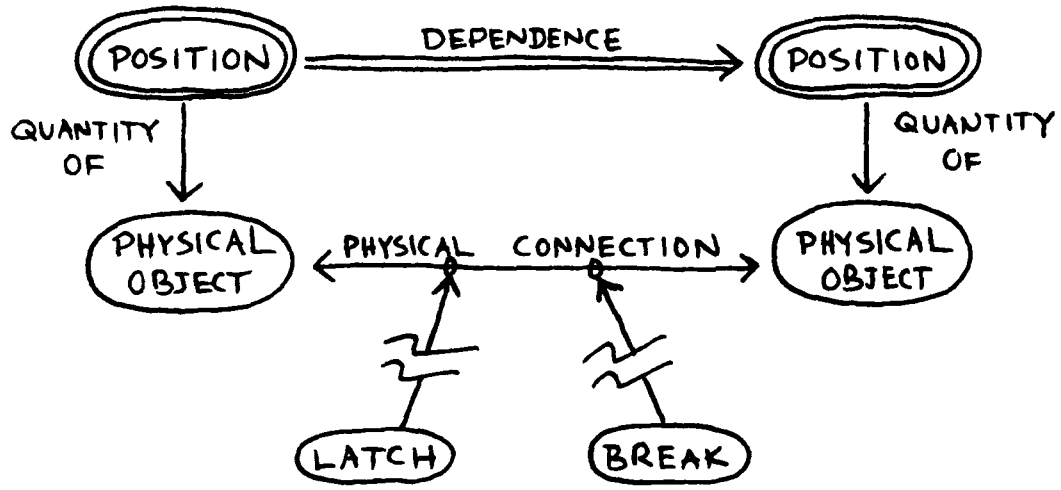


Figure 3.7: A Mechanical Coupling

### 3.3.1 A Deductive Method

The inference method which operates on the causal mechanism representation is deductive, and is concerned with recognizing proposed causal relations as instances of the *a priori* known causal mechanisms in the domain. Each causal mechanism description embodies a deductive inference rule of the form:

$$\begin{aligned}
&\exists(iq)\exists(dq)\exists(m)\neg\exists(b) \\
&\quad (IndependentQuantityType(iq) \wedge IndependentQuantityType(dq) \wedge \\
&\qquad MediumType(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq)) \\
&\qquad BarrierType(b) \wedge Along(m, b)) \\
&\Longrightarrow FunctionalDependence(iq, dq)
\end{aligned}$$

When a causal mechanism template can be fully instantiated (unified) into an observation, an instance of that causal mechanism is strongly suggested. However, the domain theory is not necessarily consistent. There may be exceptions to these inference rules. For this reason, the causal explanations generated from the domain theory are *justified*, but do not necessarily constitute *proofs* in the strong truth-preserving sense.

Sometimes the causal mechanism templates include information about the signs of proportionality relations, which can be instantiated along with the rest of mechanism descriptions. Furthermore, and of more interest, some compositions of causal mechanisms correspond to combinations of functional dependencies under addition and multiplication. The ways in which *structural* relations in causal mechanisms interact can give clues about how functional dependencies are combining additively or multiplicatively.

For example, several disjoint media adjoining the same physical object may indicate that the functional dependencies supported by those media are combining *additively*. An example is the several channels through which water may enter or leave a sink. There is the tap overhanging the basin, the drain at the bottom of the basin, and the safety drain close to the top of the basin. The net change in the level of water in the sink is the *sum* of the flows, some positive and some negative, associated with these separate media. Since the media, and the contributions, are disjoint, any one is sufficient to produce an effect. Addition is similar to disjunction in this respect. The definition of media as enabling conditions also suggests this disjunctive property.

Barriers, on the other hand, may indicate *multiplicative* contributions. Barriers can be thought of as controlling the magnitude of causal interactions along media, with a complete barrier corresponding to multiplication by zero. For example, a stopper placed over the drain in a sink will change the flow out of the drain by a multiplicative factor. When the stopper is placed securely in the drain, there is no flow. Just as addition bears similarities to disjunction, multiplication resembles conjunction, in that a single intact barrier can inhibit causation. This is also in keeping with the definition of barriers as disabling conditions.

These observations can be construed as a very simple theory of deriving function [5] from structure. I do not claim any definitive, or even terribly insightful contribution to this interesting issue. I am more interested in how deductive inference, supported by a domain theory, can complement inductive inference. There are ways in which additive and multiplicative contributions might arise other than the ones I have pointed out. For example, addition may occur via integration over time; multiplication may arise via iteration.

## 3.4 Feature Selection

The methods of causal induction outlined above are based on *elimination* of hypotheses. The inductive inferences made by the methods implicitly rely on the argument: "These are the only *known* possible causes which satisfy the constraints imposed by our assumptions about the form of the cause and our set of observations." But there is never a guarantee that the relevant causes have been considered. What is needed is some means of identifying possible causes.

This is the *feature selection* problem in learning. Inductive methods, in general, do not address it. These methods take as input a set of possible causes (or features), but they do not say how to arrive at this set. On the other hand, the domain knowledge driving the deductive approach to causal modelling does provide a partial solution to the feature selection problem. The domain knowledge is of what kinds of causal mechanisms there are. Knowledge of these mechanisms with their associated media, potential barriers, and characteristic physical objects and quantities gives the learning system guidance in knowing what is relevant, i.e., what are the possible causes. This knowledge is brought to bear during the generation of hypotheses.

It is important that this knowledge be used only as a focussing, and not an eliminative mechanism. The set of known causal mechanisms for the domain is not taken to be complete; there may be other causal mechanisms which operate in the domain. For this reason, the domain knowledge provides only a partial solution to the feature selection problem. The learning system may have to and should be able to consider causal hypotheses which cannot be justified with the current domain theory.

## 3.5 Combining Inductive and Deductive Inference

My causal modelling system utilizes both inductive and deductive inference. I have hinted at how these two inference methods interact in my system. Now I step back and consider whether there are principled ways of combining inductive and deductive inference in a learning system.

Inductive inference methods are inherently eliminative, or *falseness-preserving* [Michalski 83] while deductive methods are *truth-preserving*, or confirmative. After an inductive inference, there

[5] As in functional dependence; not in the telecological sense of function.

are more statements known to be false and anything that was false prior to the inference is still false. And conversely for deductive inference and true statements.

In the search for the correct hypothesis, inductive inference can never produce the conclusion that a particular hypothesis is the correct one, except in the unlikely event that only one hypothesis remains. Even in this situation, any such conclusion must be relative to any assumptions made, including the assumption that the correct hypothesis can be represented at all.

Deductive inference, on the other hand, can identify a correct hypothesis, as long as the domain theory is correct. Deductive inference can in fact short-circuit the need to search the hypothesis space at all.

Thus an effective strategy for combining inductive and deductive inference in a learning system is: Confirm hypotheses via deduction from the (hopefully correct) domain theory whenever possible. Fall back on induction when the domain theory does not apply (because of incompleteness). Because inductive methods are data-driven, they can continue to converge on the correct hypothesis on the basis of further observations, after the theory-driven deductive methods have nothing further to contribute.

In my learning system, the inductive method can generate hypotheses about multiple and interacting functional dependencies. The deductive method can generate justified hypotheses about individual instances of causal mechanisms. After each inductive inference, some hypotheses are eliminated. After each deductive inference, search is focussed on those causal hypotheses which have some justification. Because the domain theory consists of descriptions of individual causal mechanisms, this focussing process proceeds incrementally. If the domain theory included knowledge of compositions of causal mechanisms these focussing steps could be larger. The next section addresses exactly this shortcoming.

## 3.6 Learning New Compositions of Causal Mechanisms

The deductive method outlined above gives the learning system the ability to strongly suggest individual instances of causal mechanisms. I now show how the learning system can learn and generalize new compositions of causal mechanisms. These extensions to the domain theory of the learning system give it the ability to generate justified hypotheses involving several causal mechanisms at once. This implies a top-down approach to constructing justified causal explanations which complements the bottom-up strategy of constructing individual justified causal hypotheses.

New compositions of causal mechanisms are learned via an explanation-based, analytical learning technique [Mitchell 83, DeJong 83, Winston et al 83, Mahadevan 85]. An example in the camera domain illustrates the method. Suppose that the deductive method has strongly suggested instances of the light transmission and mechanical coupling mechanisms:

$$(Intensity(SubjectIntensity) \wedge Intensity(FilmIntensity) \wedge StraightLinePath(Path) \wedge Between(Path, Subject, Film) \wedge Opaque(Iris) \wedge Along(Iris, Path)) \Longrightarrow FunctionalDependence(SubjectIntensity \times IrisArea, FilmIntensity)$$

$$(Position(ApertureRingFStop) \wedge Position(IrisArea) \wedge PhysicalConnection(Connection) \wedge Between(Connection, ApertureRing, Iris)) \Longrightarrow FunctionalDependence(ApertureRingFStop, IrisArea)$$

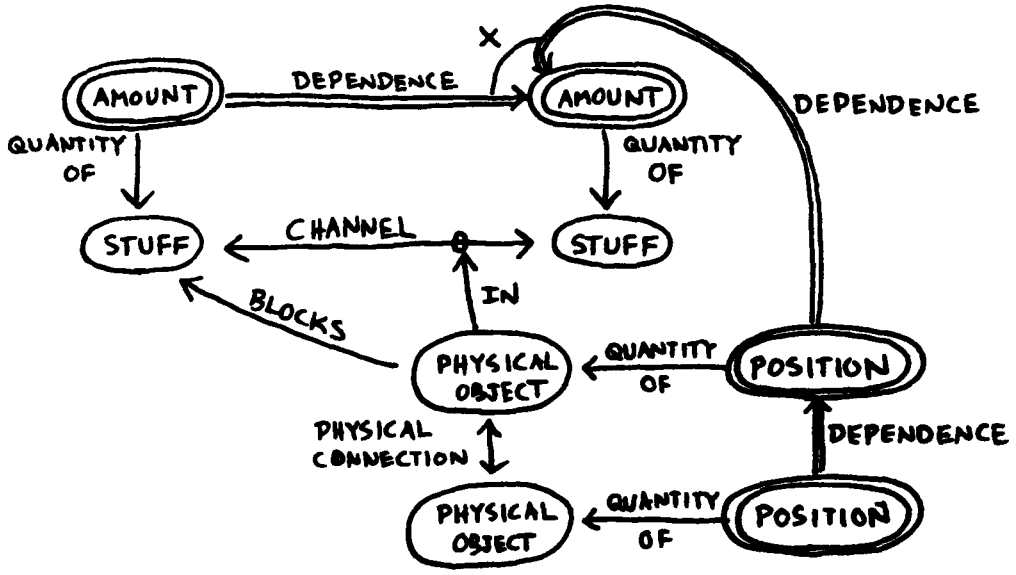


Figure 3.8: A Learned Causal Mechanism Composition

There is an interesting interaction between these causal mechanism instances. The iris, because of its opacity and its location along the straight-line path from the subject to the film, acts as a barrier to the light transmission between subject and film. The iris also participates in a mechanical coupling.

This causal model now shows how the f-stop setting can affect the film exposure – via a mechanical coupling and a flow mechanism. This multiple dependence is proposable, but not justifiable in the inductive method.

As a final step, the learning system can generalize this new composed causal mechanism by generalizing its constituent causal mechanisms. [6] Light transmission is one kind of flow. The constraint that the dependent half of the mechanical coupling also plays the role of barrier in the flow mechanism must carry through. This new mechanism corresponds to the familiar concept of a *valve*.

In summary, new compositions of causal mechanisms can be learned by inspecting the causal models generated by a combination of inductive and deductive inference. The inductive method generates possible structures of interacting causal relations (quantities and dependencies), and the deductive method justifies parts of these structures (individual causal mechanisms). Then the analytical learning method lifts out and generalizes entire justified substructures (compositions of causal mechanisms). These composed causal mechanisms can be thought of as macros or as search heuristics.

One can imagine an alternate way of "acquiring" compositions of causal mechanisms. Given the set of primitive causal mechanisms for the domain, it is possible to generate, *a priori*, all valid compositions of $2, 3, \ldots, n$ causal mechanisms.

[6] Later, when I present the set of causal mechanisms for the physical system domain, I present also a generalisation hierarchy for these mechanisms.

But there is no reason to believe that all compositions of the known causal mechanisms are interesting, useful, or even possible. The analytical learning technique uses experience as a guide in extending the domain theory. The only compositions of causal mechanisms which are generated and generalized are those which have proved relevant in the context of the construction of causal models.

## 3.7 Combining Empirical and Analytical Approaches to Learning

Recently, the field of machine learning has begun to explore knowledge-intensive analytical approaches to learning. The most successful of these methods has been *explanation-based learning.* This development is in contrast to the early stages of the field when virtually all of the results concerned general, domain-independent, empirical, inductive methods. The next logical step for the field would appear to be in exploring how these alternate approaches to learning can complement each other.

Already, some research efforts have considered how this might happen. The LEX2 system [Mitchell 83] has demonstrated how *constraint back-propagation* can extend the representation language supporting inductive learning by generating new terms in the language. Reid Smith et al [Smith et al 85] have shown how an empirical method can refine a knowledge base which might have been generated by an analytical method. I propose another way in which empirical and analytical learning methods can complement each other. In particular, I propose a way in which empirical methods can drive analytical methods.

Analytical learning methods offer what empirical methods cannot: the possibility of a guarantee that the generalizations produced by the method are correct. However, it may be not always possible to employ an analytical method, for any of several reasons: the computational cost of employing the analytical method may be deemed too great; the observation which the method is to analyze may be incomplete; the recognition of the applicability of the method is somehow inhibited. Intuitively, these are the situations in which humans, although they in principle know how to explain something, forbear from constructing the explanation because of laziness, or insufficient data, or an inappropriate censor. Often what is needed in these situations is some hint that there is indeed something which might be explained and that the effort involved in constructing the explanation or in collecting additional data is justified. Conjectures generated by empirical methods can play exactly this motivating role.

In my causal modelling system, one reason why the domain theory cannot always be brought to bear is that observations may come at a too-coarse level of perceptual granularity. Initially, observations approximate what is externally visible in the physical system at hand. Causal mechanism schemas may not be fully instantiable because the needed observations of media, quantites, etc. are not available. Assuming that the learning system has the capability of procuring observations at a finer level of granularity, – "opening up the system" – the question to ask is when should it do so?

The empirical, inductive method which proposes configurations of dependencies between quantities is less sensitive to the level of perceptual granularity as it needs only observations of quantities and can even propose quantities which are not observable. It is precisely when such empirical conjectures about dependencies between quantities exist that the learning system should be motivated to look inside the system and try to verify the presence of known causal mechanisms underlying the conjectured dependencies.

A scenario in the camera domain illustrates how the inductive method can drive the construction of justified explanations and, in turn, the explanation-based learning method. The causal modelling system can propose, purely on an empirical basis, that the film exposure appears to be proportional to the product of the brightness of the subject and the f-stop setting of the aperture ring.

$$FunctionalDependence(SubjectIntensity \times ApertureRingFStop, FilmExposure)$$

This empirical conjecture provides evidence that there are causal mechanisms at play which might be instantiable by looking inside the camera. Indeed, upon procuring observations of structure and behavior inside the camera, the learning system does instantiate several causal mechanisms to explain the dependencies which were empirically proposed (see Figure 1.2). Furthermore, the analytical learning method which identifies and generalizes compositions of causal mechanisms can now come into play as well and lift out the composition between the mechanical coupling of the aperture ring and iris, and the flow between the subject, the iris, and the film.

There are doubtless many other interesting ways in which empirical and analytical methods (and not just learning methods) can complement each other. This is likely to remain a fruitful research issue to pursue for some time.

# Chapter 4

# The Domain Theory: What Kinds of Causal Mechanisms are There?

The problem of constructing causal models is being viewed from two complementary perspectives in this research. One is a traditional inductive learning perspective which casts causal modelling as going from observations or examples (structural and behavioral descriptions of physical systems at different times) and knowledge of the general form of causal relations (compositions of dependencies between quantities) to causal models. The other perspective, embodied by the causal mechanism schema matching which tries to recognize instances of known causal mechanisms, might also be cast as a kind of learning - that of making explicit assertions which have always existed in the deductive closure. But this approach is more fruitfully interpreted as knowledge-based causal analysis – "understanding how physical systems work." The advantage of this viewpoint is the focussing of attention on the knowledge needed to make the approach successful. This knowledge is of the kinds of causal mechanisms that exist in the physical system domain.

## 4.1 The Set of Causal Mechanisms

In this section, I present the set of causal mechanisms which make up the domain theory of the causal modelling system. This part of my research is knowledge engineering. Although I have relied on my own knowledge in enumerating this set, to guard against *ad-hoc*ness I look to the field of physics to argue for the *well-foundedness* of these causal mechanisms. I show how they approximate some of the established concepts from that field.

I see no way of arguing for the completeness of this domain theory. In fact, its very incompleteness argues for a complementary inductive learning capability. However, the results of this thesis will argue for the utility of the domain theory in a number of causal modelling scenarios.

Although my system does not explicitly use *teleological* knowledge (i.e., knowledge of the purposes of things), nor does it construct explicitly teleological descriptions of physical systems, this kind of knowledge is implicit in the domain theory. For it is precisely knowledge of the kinds of causal mechanisms that exist which is exploited in the *design* of physical systems. The discovery or better understanding of such mechanisms has always led to the design of new devices. For example, a

theory of photochemical reactions in salts of silver paved the way for the design of image-recording devices.

There are three broad classes of causal mechanisms appearing in the domain theory. All describe interactions that take place in physical systems. These classes are *propagations*: causal connections between events of similar type taking place in separate objects; *transformations*: causal connections between events of dissimilar type taking place within a single object; and *field interactions*: causal connections between a field and an event taking place in an object.

All of the following causal mechanisms are described in the representation for causal mechanism given earlier.

### 4.1.1 Propagations

Two perceptually similar events occurring in different objects suggest a causal connection. This section contains descriptions of causal mechanisms which manifest in interactions of this type.

PROPAGATION

This is the generic description for the class.

$\exists(iq)\exists(dq)\exists(m)\neg\exists(b)$
$(QuantityType(iq) \wedge QuantityType(dq) \wedge$
$MediumType(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq))$
$BarrierType(b) \wedge Along(m, b))$
$\Longrightarrow FunctionalDependence(iq, dq)$

MECHANICAL COUPLING

A mechanical coupling is a transfer of motion from one object to another, mediated by a physical connection which is not broken or latched.

$\exists(iq)\exists(dq)\exists(m)\neg\exists(b)$
$(Position(iq) \wedge Position(dq) \wedge$
$PhysicalConnection(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq)) \wedge$
$((Discontinuity(b) \wedge Along(m, b)) \vee (Anchored(b) \wedge PhysicallyConnected(m, b))))$
$\Longrightarrow FunctionalDependence(iq, dq)$

FLOW

A flow is a transfer of stuff from one location to another, mediated by a channel which can carry the stuff and which is not broken or blocked. Four specific types of flows are described after this generic description.

$\exists(iq)\exists(dq)\exists(m)\neg\exists(b)$
$(Amount(iq) \wedge Amount(dq) \wedge$
$Carries(m, PhysicalObject(iq)) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq)) \wedge$
$(Discontinuity(b) \vee Blocks(b, PhysicalObject(iq))) \wedge Along(m, b))$
$\Longrightarrow FunctionalDependence(iq, dq)$

MATERIAL FLOW

A material flow is a transfer of material stuff fron one location to another, mediated by a channel which can carry the material and which is not broken or blocked.

$\exists(iq)\exists(dq)\exists(m)\neg\exists(b)$
$(MaterialAmount(iq) \wedge MaterialAmount(dq) \wedge$
$Carries(m, PhysicalObject(iq)) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq)) \wedge$
$(Discontinuity(b) \vee Blocks(b, PhysicalObject(iq))) \wedge Along(m, b))$
$\Longrightarrow FunctionalDependence(iq, dq)$

LIGHT TRANSMISSION

A light transmission is a transfer of light energy (measured by brightness) from one location to another, mediated by a straight-line path which is not blocked by an opaque object.

$\exists(iq)\exists(dq)\exists(m)\neg\exists(b)$
$(Intensity(iq) \wedge Intensity(dq) \wedge$
$StraightLinePath(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq)) \wedge$
$Opaque(b) \wedge Along(m, b))$
$\Longrightarrow FunctionalDependence(iq, dq)$

HEAT FLOW (CONDUCTION)

A heat conduction is the transfer of heat (measured by temperature) from one location to another, mediated by a path which is not broken by a vacuum (thermal conduction cannot occur across a vacuum) or blocked by a thermal insulator.

$\exists(iq)\exists(dq)\exists(m)\neg\exists(b)$
$(Temperature(iq) \wedge Temperature(dq) \wedge$
$HeatPath(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq)) \wedge$
$(Vacuum(b) \vee ThermalInsulator(b)) \wedge Along(m, b))$
$\Longrightarrow FunctionalDependence(iq, dq)$

ELECTRICITY

Finally, electricity is the transfer of electric charge from one location to another, mediated by an electrical conductor which is not physically broken or blocked by an electrical insulator.

$\exists(iq)\exists(dq)\exists(m)\neg\exists(b)$
$(ElectricCharge(iq) \wedge ElectricCharge(dq) \wedge$
$ElectricalConductor(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq)) \wedge$
$(Discontinuity(b) \vee ElectricalInsulator(b)) \wedge Along(m, b))$
$\Longrightarrow FunctionalDependence(iq, dq)$

### 4.1.2 Transformations

Transformations describe causal connections between different events occuring in the same object. The events are the externally observable aggregate manifestations of internal processes occurring at a finer level of granularity within the object. These causal mechanism descriptions do not include knowledge of media and barriers because the internal processes are assumed to be unavailable for scrutiny.

Transformations across separate objects are explained by a combination of a transformation within one of the objects and a propagation from or to the other object.

TRANSFORMATION

This is the generic description for the class. Transformation descriptions are impoverished relative to propagation descriptions.

$\exists(iq)\exists(dq)$
$(IndependentQuantityType(iq) \wedge DependentQuantityType(dq))$
$\Longrightarrow FunctionalDependence(iq, dq)$

ELECTROMECHANICAL

In the electromechanical transformation, electrical energy at an object is transformed into motion.

$\exists(iq)\exists(dq)$
$(ElectricCharge(iq) \wedge Position(dq))$
$\Longrightarrow FunctionalDependence(iq, dq)$

ELECTROPHOTIC

In the electrophotic transformation, electrical energy at an object is transformed into light energy.

$\exists(iq)\exists(dq)$
$(ElectricCharge(iq) \wedge Intensity(dq))$
$\Longrightarrow FunctionalDependence(iq, dq)$

ELECTROTHERMAL

In the electrothermal transformation, electrical energy at an object is transformed into thermal energy.

$\exists(iq)\exists(dq)$
$(ElectricCharge(iq) \wedge Temperature(dq))$
$\Longrightarrow FunctionalDependence(iq, dq)$

PHOTOCHEMICAL

In the photochemical transformation, light energy at an object is transformed into a change of appearance (e.g. color).

$\exists(iq)\exists(dq)$
$(Intensity(iq) \wedge Appearance(dq))$
$\Longrightarrow FunctionalDependence(iq, dq)$

THERMOCHEMICAL

In the thermochemical transformation, thermal energy at an object is transformed into a change of appearance. This mechanism describes burned objects.

$\exists(iq)\exists(dq)$
$(Temperature(iq) \wedge Appearance(dq))$
$\Longrightarrow FunctionalDependence(iq, dq)$

EXPANSION

Finally, in the expansion transformation, thermal energy at an object is transformed into a change in size or position.

$\exists(iq)\exists(dq)$
$(Temperature(iq) \wedge Position(dq))$
$\Longrightarrow FunctionalDependence(iq, dq)$

### 4.1.3 Field Interactions

Field interactions are the most bizarre of the causal mechanisms included in the domain theory of the causal modelling system. The strangeness comes from the fact that the field source does not *do* anything observable which can be correlated with the motion of an object interacting with the field. The "cause" is really the relative position of the affected object within the field.

FIELD INTERACTION

This is the generic description for the class.

$$\exists(fs)\exists(dq)$$
$$(FieldSource(fs) \wedge Position(dq) \wedge$$
$$FieldProperty(PhysicalObject(dq)))$$
$$\Longrightarrow FunctionalDependence(PositionOf(fs) - dq, dq)$$

GRAVITATION

Any object with mass interacts with a gravitational field.

$$\exists(fs)\exists(dq)$$
$$(GravitationalFieldSource(fs) \wedge Position(dq) \wedge$$
$$Massed(PhysicalObject(dq)))$$
$$\Longrightarrow FunctionalDependence(PositionOf(fs) - dq, dq)$$

MAGNETISM

Any magnetic or electrically charged object interacts with a magnetic field.

$$\exists(fs)\exists(dq)$$
$$(MagneticFieldSource(fs) \wedge Position(dq) \wedge$$
$$(Magnetic(PhysicalObject(dq)) \vee Charged(PhysicalObject(dq))))$$
$$\Longrightarrow FunctionalDependence(PositionOf(fs) - dq, dq)$$

ELECTRIC FIELD INTERACTION

Any electrically charged or magnetic object also interacts with an electrical field.

$$\exists(fs)\exists(dq)$$
$$(ElectricFieldSource(fs) \wedge Position(dq) \wedge$$
$$(Charged(PhysicalObject(dq)) \vee Magnetic(PhysicalObject(dq))))$$
$$\Longrightarrow FunctionalDependence(PositionOf(fs) - dq, dq)$$

### 4.1.4 The Generalization Hierarchy

Knowledge of the valid generalizations within a domain is clearly knowledge which can drive a learning system. The causal mechanism types listed above fit into a generalization hierarchy as follows:

This knowledge is used by the explanation-based learning method described earlier to generalize new compositions of causal mechanisms found in the course of constructing causal models.

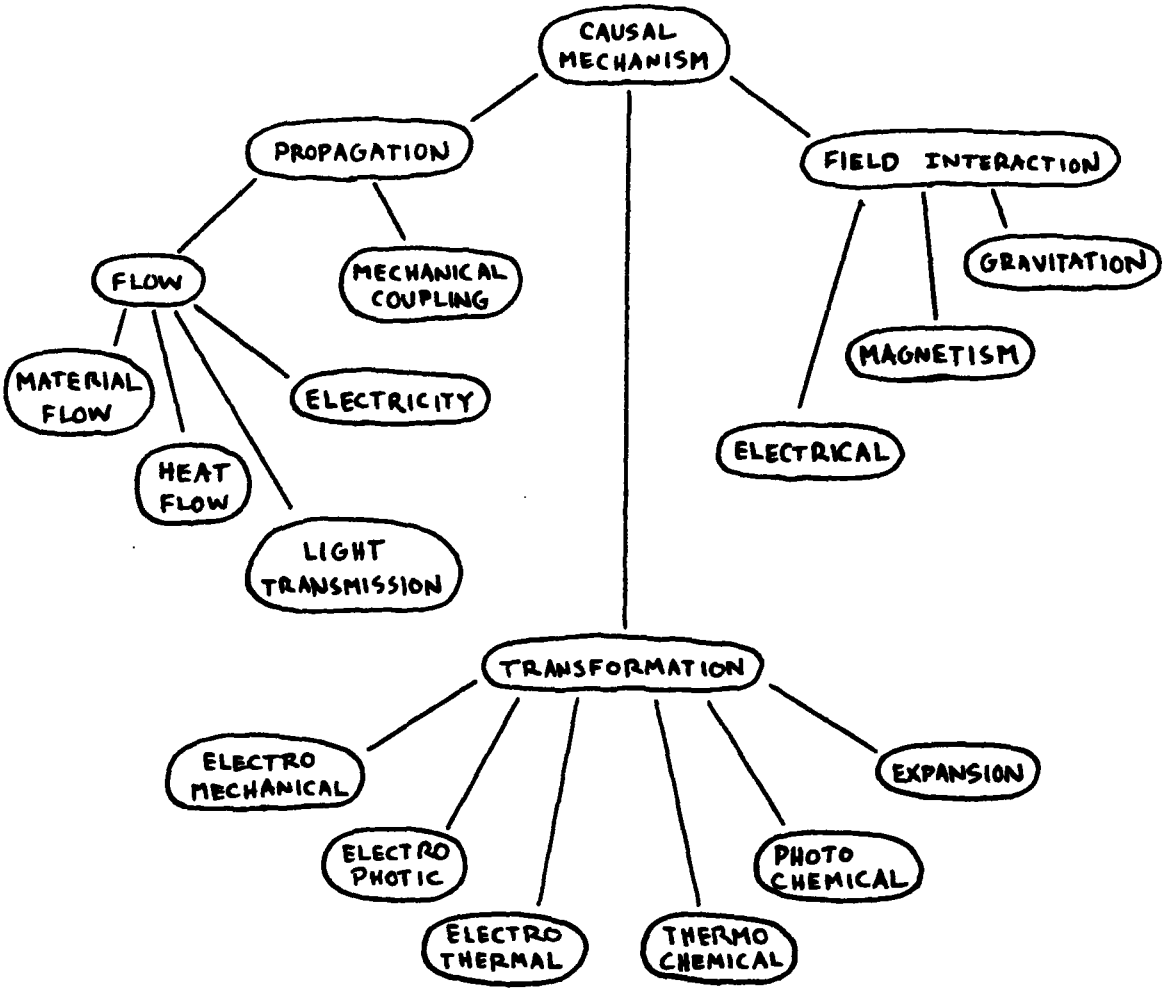


Figure 4.1: The Generalization Hierarchy for Causal Mechanisms

## 4.2 Looking to Physics: Conservation

In this section, I argue for the well-foundedness of the causal mechanism descriptions which comprise my causal modelling system's domain theory. I relate these descriptions to the *conservation* concept from physics.

Conservation is one of the most important concepts in physics. In its most abstract form this concept states that the amount of any quantity associated with any isolated closed system must remain constant.

$$\sum q_i = c$$

The distribution of the total amount of a quantity within a closed system may change. In other words, fractions of the total amount of a quantity may propagate to different locations within the system. In addition, a quantity may manifest in different forms and fractions of the total amount of a quantity may change from one form to another within the system. But at no time does the total amount of a conserved quantity within the system change.

A corrollary of the statement of conservation states that when some amount of a conserved quantity is introduced (or removed) from an otherwise closed system, the total amount of that quantity in the system changes by exactly that amount. [1]

$$\sum q_i + \Delta q = c + \Delta q$$

Again, the introduced or removed amount may be propagated or it may undergo transformations. But at no time is some fraction of the total amount of a conserved quantity in a closed system consumed or created.

I submit that all of the causal mechanisms listed above can be interpreted in terms of conserved quantities in closed systems. An effect is never an isolated self-generative incident; it can always be explained in terms of a propagation or a transformation within a closed system, or ultimately, in terms of a propagation or transformation originating in an external input to a system.

The causal mechanisms called propagations are straightforwardly interpreted as reflections of conservation laws. The conserved quantities are the ones which appear in the causal mechanism descriptions as both cause and effect. The mechanical coupling mechanism describes co-occuring changes in position. This corresponds to conservation of momentum. Similarly, material flow is conservation of mass; light transmission and heat flow are conservation of (electromagnetic) energy; electricity is conservation of charge.

The causal mechanisms involving transformations also reflect conservation laws. However, the quantities being conserved take on different perceptual forms, which makes the effect difficult to recognize as another manifestation of the cause. For example, in expansion, a change in the kinetic energy of molecules in an object, manifested as a change in temperature, is also manifested as a change in the physical size of the object. In the photochemical reaction, electromagnetic energy impinging on an object is transformed into an increase in the kinetic energy of the object's molecules. The agitated molecules rearrange various chemical bonds – a process which manifests as a change

[1] In the case of conservation of energy, for example, this introduced/removed amount is characterized as work done on/by the system.

in appearance of the object. The other transformation mechanisms are also interpretable as conservations of energy in its various forms. Thus transformations really are propagations; the quantity being transferred is energy.

In field interactions, potential energy is interchanged with kinetic energy. The strangeness of these mechanisms (and perhaps one source of difficulty for scientists who tried to understand them) probably arises from the fact that potential energy is not observable as a local property of an object, rather it is related to the global position of an object within a field. Hence a cause or effect involving potential energy is not easily identified. In any event, field interactions are also a kind of propagation governed by a conservation law, once again conservation of energy.

Of course, very few designed physical systems which "implement" these conservation laws are truly closed, as the strict interpretation of conservation requires. In other words, the amount of a quantity at the effect after a propagation from a cause may not be totally conserved. For example, some of the kinetic energy of a mechanical coupling may be dissipated as thermal radiation; a material transport system (e.g. plumbing) may have leaks; some of the energy of motion of a rock falling in a gravitational field may be dissipated as thermal radiation via air resistance, etc.

But this is not the point of this conservation-based interpretation of causality. Rather, the point is to demonstrate that the causal mechanisms which make up the causal modelling system's domain theory really are *causal*, in keeping with the philosophers' notion of *necessity* in causality. For conservation tells us that indeed the effect necessarily follows from the cause, is produced by the cause, emerges from the cause, for the effect – propagations and transformations notwithstanding – is always the same *stuff* as the cause.

## 4.3 Indexing the Mechanisms

Earlier, I argued that a deductive inference method can constrain search through an hypothesis space. Once a strongly justified hypothesis is found, there may be no need to consider alternate hypotheses which are consistent with observations, and which an inductive method could not eliminate. However, a deductive inference method substitutes a search for a structure of deductive inferences (a proof or explanation) in place of a search for hypotheses.

A search for hypotheses can be heuristically constrained by making assumptions about the form of the correct hypothesis. One way to heuristically constrain the search for an explanation is to index inference rules under summaries of their contents. If an explanation is desired for something which matches the summary of a rule, this provides motivation for an attempt to fully instantiate the rule. Indexing can be even more useful if *macros* (compositions of several inference rules) exist. Matching the summary of a macro can motivate the instantiation of an entire network of inferences.

In my causal modelling system, explanations are instantiated causal mechanism schemata. Primitive causal mechanism schemata are indexed under the types of quantities appearing in the functional dependencies for which they can provide justified explanations. As compositions of causal mechanisms are acquired, they are indexed for all the functional dependencies they can explain.

For example, the mechanical coupling mechanism is indexed under *position* and *position*. The composed flow and mechanical coupling mechanism (the valve mechanism) is indexed under *amount* and *amount*, *position* and *position*, *position* and *amount*.

Causal mechanism schemata are summarized for indexing in terms of quantity types precisely because the empirical conjectures about causal relations generated by the inductive method are in

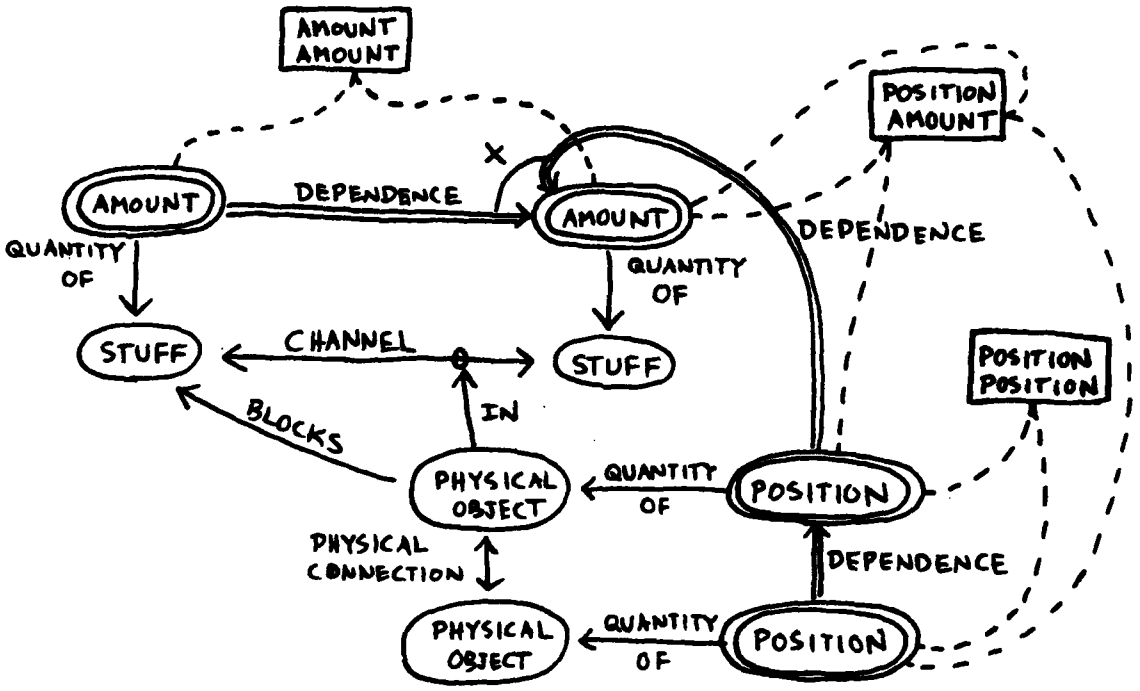


Figure 4.2: Indexes for the Valve Mechanism

terms of dependencies between quantities. Empirical conjectures may be *abstractions* of the more detailed causal explanations embodied by the mechanism descriptions. It is this abstraction relation which is captured by indexing and presented as a search heuristic.

# Chapter 5

# Levels of Abstraction

Multiple levels of abstraction impinge on the causal modelling problem in two ways: Physical systems can be observed at different levels of *perceptual* resolution; and causal *explanations* of observations can be constructed at different levels of detail. The interesting issue concerning multiple levels of abstraction is one of *control*: To what level should perceptions and explanations go?

In the rest of this section, I elucidate the following arguments towards a principled solution to this issue. In general, the tasks of controlling levels of perception and levels of explanation are not independent. The level of explanation is determined in part by what needs to be explained, i.e., by the motivating causal reasoning task which provides the context for causal modelling. Some explanations intrinsically require more detail than others. The required level of explanation in turn determines the required level of perception. Certain observations must be collected to instantiate an explanation. Conversely, difficulties in obtaining the required observations can affect the level of explanation. In the worst case, it is not possible to contruct an explanation because the required observations are unobtainable. In other cases, it is possible to construct an alternate explanation at a different level of abstraction.

## 5.1 Levels of Explanation in the Domain Theory

The causal mechanism descriptions which make up the causal modelling system's domain theory become explanations for observations when they are instantiated into observations. Only a single level of explanation in these causal mechanism descriptions has been discussed thus far. Recall the description for propagations:

$$\begin{aligned}&\exists(iq)\exists(dq)\exists(m)\neg\exists(b)\\&\quad(QuantityType(iq)\wedge QuantityType(dq)\wedge\\&\quad MediumType(m)\wedge Between(m,PhysicalObject(iq),PhysicalObject(dq))\\&\quad BarrierType(b)\wedge Along(m,b))\\&\Longrightarrow FunctionalDependence(iq,dq)\end{aligned}$$

Actually, each mechanism description implicitly contains many layered explanations, each approximating the one at the next level of detail, until the description bottoms out.

The most abstracted causal explanation derivable from a propagation description is simply that two quantities are both of the same type.

$$\exists(iq)\exists(dq)$$
$$(QuantityType(iq) \wedge QuantityType(dq) \wedge$$
$$\Longrightarrow FunctionalDependence(iq, dq)$$

For example, an explanation for why bread placed in a toaster gets hot can be simply that the coils in a toaster also get hot.

This explanation is an abstraction of one which includes the structural link between the physical objects of the quantities.

$$\exists(iq)\exists(dq)\exists(m)$$
$$(QuantityType(iq) \wedge QuantityType(dq) \wedge$$
$$MediumType(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq))$$
$$\Longrightarrow FunctionalDependence(iq, dq)$$

Now the explanation for why bread placed in a toaster gets hot can include the thermally conducting medium of the air between the coils and the bread.

This explanation in turn approximates one which states that there must be no barriers which can decouple a structural link.

$$\exists(iq)\exists(dq)\exists(m)\neg\exists(b)$$
$$(QuantityType(iq) \wedge QuantityType(dq) \wedge$$
$$MediumType(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq))$$
$$BarrierType(b) \wedge Along(m, b))$$
$$\Longrightarrow FunctionalDependence(iq, dq)$$

The toaster explanation now can be elaborated to state that there must be no insulating material along the path from the coils to the bread.

Finally, this description of a barrier can be fleshed out to state that in general the effectiveness of a barrier depends on how much of the medium it blocks. Notice that this level of description reveals a dependence which does not appear at any of the higher levels.

$$\exists(iq)\exists(dq)\exists(m)\exists(b)$$
$$(QuantityType(iq) \wedge QuantityType(dq) \wedge$$
$$MediumType(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq))$$
$$BarrierType(b) \wedge Along(m, b))$$
$$\Longrightarrow FunctionalDependence(iq \times CrossSectionalAreaOf(b, m), dq)$$

The statement about insulators in toasters now can be generalized to: Any thermal insulator reduces the change in temperature of the bread due to a change in temperature of the coils by an amount proportional to the amount of cross-sectional area along the path from the coils to the bread blocked by the insulator.

## 5.2 A Continuum of Explanations from Empirical to Proven

There is an even more abstract explanation in this hierarchy of explanations than any of those derivable from the domain theory. This type of explanation has been described previously as *associative*;

it merely associates one event, the cause, with another, the effect, without offering any justified reason why the effect should follow from the cause.

$$\exists(iq)\exists(dq)$$
$$(ValueAt(iq,t1) \wedge ValueAt(dq,t2) \wedge \geq (t1,t2))$$
$$\Longrightarrow FunctionalDependence(iq,dq)$$

Associative causal explanations exploit the regularity and direction aspects of causality, but not the mechanism aspect. It is entirely consistent that these associative explanations are arrived at via inductive inference. An inductive inference is never justified; it can have only an empirical basis.

The other explanations in this hierarchy, derived from the domain theory, are all *mechanistic* to varying degrees. They all attempt to offer justified causal explanations in terms of known mechanisms underlying causal interactions. These explanations make use of deductive inference and are proof-like, but they should not be construed as proofs in the full truth-preserving sense. Each mechanistic explanation has exceptions, namely the ones specified at lower levels in the explanation hierarchy. In other words, each explanation is a *non-truth-preserving approximation* of explanations found lower in the hierarchy. And there is no reason to believe that further exceptions do not exist below where the explanation hierarchy bottoms out.

Rather than a simple dichotomy between associative and mechanistic causal explanations, or between inductive and deductive inference, the explanation hierarchy represents a continuum of explanations from ones with only an empirical basis, to ones with less and less of an empirical flavor and greater and greater justification. Because the domain theory is not strong enough to provide real proofs grounded in irrefutable axioms, even the most justified explanations in this hierarchy still have a touch of empiricism, in that they are not full accountings of all situations in which they apply; they are not universal statements.

One kind of abstraction which is different from the *approximations* discussed above is *aggregation*. This is the kind of abstraction in which many primitives at one level are subsumed under a single primitive at the next. Furthermore, the vocabulary of primitives at different levels may change. Some of the causal mechanism descriptions characterized as *transformations* involve aggregations. For example, the expansion schema describes a causal relation between two aggregate properties of physical objects - temperature and size. Both of these macroscopic properties have microscopic explanations in terms of the motions of molecules. These low-level explanations do not appear in the expansion schema because I do not have a way of representing how many objects and mechanisms at one level can be subsumed under a single object and quantity at the next level. [1] I represent aggregations only at the aggregate level. The following discussions only relate to approximation abstractions.

## 5.3 Some Properties of Explanations

Before discussing the issue of how to control the level of explanation and perception, I introduce some properties of the assertions which appear in causal mechanism descriptions, and of explanations derived from them.

[1] See [Weld 84] for a solution to this problem.

### 5.3.1 Instantiability

An assertion is *instantiable* if and only if it is easily computable or, better yet, directly unifiable from the perceptions in an observation input to the causal modelling system. For example, the assertion:

$$Between(m, PhysicalObject(iq), PhysicalObject(dq))$$

from the material amount flow schema is instantiable from the following set of perceptions

$In(Pipe1, Water1)$
$In(Pipe1, Water2)$
$QuantityOf(Water1, Amount1)$
$QuantityOf(Water2, Amount2)$

with the following unifications

$$(m/Pipe1)(iq/Amount1)(dq/Amount2)$$

and some simple inference rules concerning the definition of *Between* and the inverse relation between $QuantityOf$ and *PhysicalObject*.

Instantiability is essentially an *operationality* criterion.

An explanation is instantiable if and only if all of its assertions are instantiable.

### 5.3.2 Observability

An assertion is *observable* if and only if there is an observation in which the term is instantiable. The same assertion

$$Between(m, PhysicalObject(iq), PhysicalObject(dq))$$

from the heat flow schema is unobservable in the toaster domain because perceptions needed to instantiate it, such as

$Contiguous(Air1, Coils1)$
$Contiguous(Air1, Bread1)$

will never appear in an observation. Observability is a statement about the limitations of some assumed perception equipment.

An explanation is observable if and only if there is an observation in which all of its assertions are instantiable.

### 5.3.3 Consistency

An explanation is *consistent* if and only if the inferences which it supports are corroborated in all observations in which the explanation is instantiated.

Thus an explanation which includes media but not barriers can be inconsistent with a set of observations which includes at least one where a barrier is intact. If the explanation is instantiable for the observation involving the intact barrier, it would support the conclusion that the causal mechanism whose medium was instantiated is active, i.e., that the indicated functional dependencies will hold. This would be the wrong conclusion.

An explanation which includes the appropriate barriers would be consistent with the same set of observations. In the case where the barrier was in place, it would support the inference that the indicated functional dependencies will not hold.

## 5.4 Controlling the Levels of Explanation and Perception

The issue "To what level should explanation and perception go?" can be clarified by first determining what kinds of causal descriptions, independent of level, the causal modelling system should produce. I want my causal modelling system to produce *mechanistic*, *consistent* causal descriptions.

The argument for mechanistic descriptions is twofold: Mechanistic descriptions have some justification; they are an attempt to reduce the empirical basis of causal explanations. From the philosphers' viewpoint, mechanistic descriptions begin to get at necessity in causation; they suggest why the effect emerges from the cause; is produced by the cause.

The other argument for mechanistic descriptions has to do with differences between inductive and deductive inference. Associationist descriptions are arrived at via inductive inference. Inductive inference involves exponential search through a hypothesis space. Deductive inference can short-circuit this search by providing "proofs", i.e., justified explanations for particular hypotheses. This focussing must be traded off against the amount of search required to generate explanations. This kind of search is kept manageable by indexing causal mechanism descriptions – the templates for explanation – under the types of causal hypotheses they can explain.

The argument for consistent explanations is in the context of the causal reasoning tasks which causal models are intended to support. The utility of learned causal models is severely restricted if they do not incorporate consistent causal explanations/descriptions, because any inference made by a reasoning program operating on an inconsistent causal model is suspect.

Given that the causal modelling system should produce mechanistic and consistent causal descriptions, the question "To what level should explanation and perception go?" can be rephrased to "What level of explanation and perception is required to produce mechanistic and consistent causal descriptions?"

### 5.4.1 Controlling the Level of Explanation

For the moment, I ignore the perception issue and consider only the level of explanation required to produce mechanistic and consistent causal descriptions.

- What level of explanation is required to produce mechanistic causal descriptions?

There is a glib answer to this question. All levels of explanation in the domain theory are mechanistic. Thus any explanation derived from the domain theory is sufficient to produce mechanistic causal descriptions.

But this glib answer ignores the fact that different causal descriptions/explanations, all mechanistic, require different amounts of detail. For example, the most detailed flow schema

$$\begin{aligned}&\exists(iq)\exists(dq)\exists(m)\neg\exists(b)\\&\quad(Amount(iq)\wedge Amount(dq)\wedge\\&\quad Carries(m, PhysicalObject(iq))\wedge Between(m, PhysicalObject(iq), PhysicalObject(dq))\wedge\\&\quad(Discontinuity(b)\vee Blocks(b, PhysicalObject(iq)))\wedge Along(m,b))\\&\Longrightarrow FunctionalDependence(iq\times CrossSectionalAreaOf(b,m), dq)\end{aligned}$$

can be used to generate an explanation for a causal relation between two amount quantities, or an explanation for a causal relation between an area quantity and an amount quantity.

A sufficient explanation, still mechanistic, for the amount-amount proportionality could have been generated from a higher-level abstraction of this schema, such as:

$$\begin{aligned}&\exists(iq)\exists(dq)\exists(m)\\&\quad(Amount(iq)\wedge Amount(dq)\wedge\\&\quad Carries(m, PhysicalObject(iq))\wedge Between(m, PhysicalObject(iq), PhysicalObject(dq))\wedge\\&\Longrightarrow FunctionalDependence(iq, dq)\end{aligned}$$

However, the area-amount proportionality requires the more detailed explanation. The area quantity cannot be instantiated until both a medium and a barrier are instantiated. From another viewpoint, a medium can be described in the absence of barriers, but a barrier without the context of a medium makes no sense.

To summarize, many causal mechanism descriptions can explain more than one functional dependence. [2] For each explainable functional dependence, there is a minimal context which must be instantiated; this corresponds to the required level of explanation.

- What level of explanation is required to produce consistent causal descriptions?

There is no *a priori* answer to this question. The consistency of a description is relative to the observations into which it has been instantiated. For example, a causal explanation for water flowing into a sink in terms of a pipe carrying water from somewhere else is perfectly consistent until the faucet is turned off. Then this description becomes inconsistent because it still supports the prediction that water will flow into the sink. However, a causal explanation at a greater level of detail which incorporates knowledge of barriers remains consistent. It explains both why water flows when the faucet is on and why it does not flow when the faucet is off.

Thus explanations may have to become more detailed to become more general, i.e., to be consistent with more and more observations. The determination of the level of explanation required to maintain consistency is failure-driven. Only if the domain theory itself is known *a priori* to be consistent, can the explanations derived from it also be declared *a priori* to be consistent. This is not the case for the domain theory of the causal modelling system, or any domain theory which contains layered approximations. Approximations imply inconsistency.

[2] This is always true of composed causal mechanism descriptions.

### 5.4.2 Controlling the Level of Perception

An explanation is of no use until it has been instantiated into an observation. The level of explanation required to produce a mechanistic and consistent causal description can be derived directly from a domain theory, but whether or not an explanation can be instantiated depends ultimately on the limitations of the perceptual equipment which procure observations. There are two reasons why an explanation might not be instantiable in an observation. One is that the required perceptions can only be obtained at a finer level of perceptual granularity. The other is that the explanation involves terms which are intrinsically unobservable.

In this section, I explore how limitations of perception interact with the level of explanation, in some cases limiting the attainable level of explanation, in others suggesting that a more detailed explanation would be useful.

When an explanation is uninstantiable, this does not necessarily imply that the explanation is unobservable. It may be that the perceptions which would instantiate the explanation are available at the next level of perceptual granularity. In other words, if the physical system in question were "opened up", the revealed perceptions might be sufficient to instantiate the explanation.

For example, explaining a light bulb in terms of a source of electricity and a wire conducting that electricity might require looking inside the wall to observe the wire.

This suggests a heuristic for controlling the level of perception in support of the instantiation of a desired explanation. This heuristic might be called the *Look-Inside* heuristic.

> IF an explanation cannot be instantiated at the current level of perception,
> THEN obtain an observation at the next level of perception.

Presumably, there is always some finest level of perception beyond which the *Look-Inside* heuristic cannot be applied. This finest level of perception limits the attainable level of explanation. Explanations which require perceptions beyond this level cannot be instantiated. It is possible that no mechanistic explanations for a causal relation can be instantiated, in which case the causal modelling system has no choice but to employ inductive inference and generate purely empirical explanations. In the absolutely worst case, even the observations of quantities required to support inductive inference may be unavailable, so that no causal modelling is possible.

It may be the case that an explanation is uninstantiable because it involves an assertion which is intrinsically unobservable. Examples of unobservable [3] terms might be the conducting medium of the air inside a toaster supporting heat flow or the vacuum which is a sufficient channel for light (or any electromagnetic) transmission.

There is no level of perceptual granularity at which an intrinsically unobservable assertion can be instantiated. An explanation which requires an unobservable assertion also must be declared unobservable.

But this does not mean that a partially instantiated explanation is useless, particularly since all explanations derived from the domain theory have some empirical basis. A partially instantiated explanation can be checked for consistency in the same way as a fully instantiated explanation – by corroborating the predictions it supports for other observations into which it can be (partially)

[3] Observability is always relative to some assumed perception equipment.

instantiated. If the partially instantiated explanation is found to be consistent, this is empirical evidence that the unobservable assertion does in fact hold. Empirical evidence for the explanation has to be substituted for the justification that the instantiated assertion would have provided.

The greater empirical nature of an explanation enforced by an unobservable assertion can be countered by switching to an explanation that is intrinsically more justified, i.e., a more detailed explanation in the explanation hierarchy which applies to more situations. If the explanation at the more detailed level is found to be consistent, this empirical evidence makes for a stronger case that the unobservable assertion holds and is supporting the causal mechanism. This empirical evidence, coupled with the stronger intrinsic justification of the more detailed explanation, counters the loss of justification due to the unobservable term.

As an example, the following explanation from the light transmission schema might be chosen for instantiation as part of the construction of a causal model for a camera.

$$\begin{aligned}&\exists(iq)\exists(dq)\exists(m)\\&\quad(Intensity(iq) \wedge Intensity(dq) \wedge\\&\quad StraightLinePath(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq)) \wedge\\&\Longrightarrow FunctionalDependence(iq, dq)\end{aligned}$$

Assume that the term $StraightLinePath(m)$ in this explanation is unobservable. Further observations in which the expected proportionality between intensities holds can provide empirical evidence for the existence of the light path. Better yet, a more detailed, more justified, more general explanation can be partially instantiated and tested for consistency empirically; the greater intrisic justification of the more detailed explanation counters the loss of justification due to the unobservable light path.

For the camera, the barrier which appears in the next most detailed explanation can be instantiated (e.g. a lens cap) and the expected disablement of the light transmission mechanism can be corroborated.

$$\begin{aligned}&\exists(iq)\exists(dq)\exists(m)\neg\exists(b)\\&\quad(Intensity(iq) \wedge Intensity(dq) \wedge\\&\quad StraightLinePath(m) \wedge Between(m, PhysicalObject(iq), PhysicalObject(dq)) \wedge\\&\quad Opaque(b) \wedge Along(m, b))\\&\Longrightarrow FunctionalDependence(iq, dq)\end{aligned}$$

A partially instantiated explanation surely cannot guarantee truth-preserving inferences, but the inferences it does support are no less useful than purely inductive inferences or the other non-truth-preserving inferences supported by the layered approximations in the domain theory.

## 5.5 Summary

In summary of this chapter, the insights gained into the issue of "To what level should explanation and perception go?" appear below:

- The domain theory embodies a hierarchy of explanations from purely empirical to more and more justified.
- There is a minimal context which must be instantiated for any explanation; this context corresponds to some level in the explanation hierarchy.

- A more detailed explanation may be required to maintain consistency.
- An uninstantiable explanation may be instantiable in an observation at a finer level of perceptual resolution.
- An unobservable term implies loss of justification which can be countered with a more detailed, less approximate explanation.

# Chapter 6

# Learning from Experiments

The ability to design experiments changes a learning system from a passive one to an active one. The overall goal of the causal modelling system is to converge on a single, strongly justified hypothesis of what causal relations exist in a given physical system. This process can be accelerated by deliberately arranging for observations which provide further justification for leading hypotheses or distinguish competing hypotheses.

There is a recurrent theme of two flavors of knowledge in the causal modelling system: one kind of knowledge is about compositions of functional dependencies between quantities; this kind of knowledge supports inductive, entirely empirical reasoning which only can eliminate hypotheses. The other kind of knowledge is about the kinds of causal mechanisms which exist in the physical system domain; it supports reasoning which is more analytic, deductive in form, and leads to justified hypotheses with less, but some empirical flavor. Both of these kinds of knowledge can support the design of experiments to accelerate the causal modelling process. Not surprisingly, the knowledge which supports inductive reasoning can be used only to design experiments to eliminate hypotheses; the knowledge which supports analytic reasoning can be used to design experiments to both further justify hypotheses and to distinguish hypotheses.

In general, learning from an experiment involves designing an experiment which generates certain expectations or predictions, and then adjusting hypotheses according to the results of the experiment. The ideal experiment intended to distinguish hypotheses is one which generates different and unambiguous predictions for all competing hypotheses. The results of such an experiment isolate exactly one hypothesis as the correct one.

More formally, an experiment partitions hypotheses into equivalence classes corresponding to different predictions. The ideal experiment is one where each equivalence class contains a different, and exactly one, hypothesis. In practice, these ideal experiments are extremely difficult to come by.

Experiment design involves instantiating a situation and mapping hypotheses into equivalence classes based on the predictions they make for that situation. The hard part of experiment design is that the situation should be chosen so that hypotheses are evenly distributed over as many equivalence classes as possible.

## 6.1 Experiments Based on Knowledge of Functional Dependencies

The inductive learning method used by the causal modelling system generates hypotheses about causal relations in the form of possible compositions of functional dependencies under negation, multiplication, and addition. An example of one of these hypotheses is the composed functional dependence $E = (-A) \times (B + C)$. Experiment design involves assigning values to the independent quantities in hypotheses and evaluating for predictions of the values of dependent quantities.

One way to design an experiment to distinguish hypotheses is to enumerate situations (value assignments) until one is found in which the set of hypotheses show some divergence in their predictions for that situation.

For example, say we want to design an experiment to distinguish the hypotheses $E = (-A) \times (B + C)$ and $E = A \times (B \times C)$. We can simply enumerate assignments from the set of possible values $\{0+-\}$ to the variables $A$, $B$, and $C$ until one is found in which the two hypotheses generate different predictions for the value of $E$.

| SITUATION | | | PREDICTION OF | PREDICTION OF |
|---|---|---|---|---|
| $A$ | $B$ | $C$ | $(-A) \times (B + C)$ | $A \times (B \times C)$ |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | + | 0 | 0 |
| 0 | 0 | − | 0 | 0 |
| 0 | + | 0 | 0 | 0 |
| 0 | + | + | 0 | 0 |
| 0 | + | − | 0 | 0 |
| 0 | − | 0 | 0 | 0 |
| 0 | − | + | 0 | 0 |
| 0 | − | − | 0 | 0 |
| + | 0 | 0 | 0 | 0 |
| + | 0 | + | + | 0 |

In the worst case, the search for a useful experiment is clearly exponential in the number of independent quantities which appear in hypotheses. [1]

This search can be constrained heuristically by making use of knowledge about constraints on the ways functional dependencies combine under the composition operators. For example, knowledge that the result of multiplication by 0 is always 0 could have prevented the experiment designer from considering any assignments in which $A$, which is a top-level multiplicand in both hypotheses above, is 0. Also, knowledge that sums and products diverge when one operand is zero and one non-zero could have focussed the experiment designer to assignments where exactly one of $B$ and $C$ is zero. This kind of heuristic knowledge should be made available to the experiment designer.

### 6.1.1 Ambiguity in Experiments

The design of experiments can be complicated by the fact that there may be several possible outcomes for a situation. For example, the sum of a positive and a negative value may be positive, negative,

[1] The number of predictions that may have to be computed in the worst case is $h \times p^v$ where $p$ is the number of possible predictions, $v$ is the number of independent quantities, or variables, and $h$ is the number of hypotheses to be distinguished. The experiment design problem is clearly in *NP*.

or zero. Some indeterminism is unavoidable with qualitative values and can lead to ambiguity in the results of experiments.

Three types of experiments of varying degrees of ambiguity can be distinguished:

- An *unambiguous* experiment is one where the sets of predictions for all hypotheses are mutually disjoint. The results of an unambiguous experiment isolate exactly one hypothesis as the correct one.

The experiment corresponding to the last line in the table above is unambiguous.

- A *potentially ambiguous* experiment is one where there is a pairwise intersection among the sets of predictions for all hypotheses which is non-null. The results of a potentially ambiguous experiment may eliminate any number of hypotheses from none to all but one.

The following experiment, not shown above, is potentially ambiguous.

| SITUATION | | | PREDICTION OF | PREDICTION OF |
|---|---|---|---|---|
| $A$ | $B$ | $C$ | $(-A) \times (B + C)$ | $A \times (B \times C)$ |
| + | + | − | 0 + − | − |

If the outcome of the experiment is 0 or +, then $(-A) \times (B + C)$ must be the correct hypothesis; if the outcome is −, there is no conclusion.

- A *hopelessly ambiguous* experiment is one where the sets of predictions for all hypotheses are equal. The results of a hopelessly ambiguous experiment cannot eliminate any hypotheses.

All the experiments in the table above except the one corresponding to the last line are hopelessly ambiguous.

The built-in indeterminism of the qualitative calculi is not the only source of ambiguity in experiments. The pigeonhole principle reveals another source. Whenever the number of hypotheses is greater than the number of possible predictions, there must be ambiguity.

This observation exposes a rather severe limitation on experiment design imposed by the qualitative representation for the values of quantities. Given that there are only three possible predictions for the values of quantities under the qualitative representation, any experiment involving four or more hypotheses is potentially ambiguous.

This points up once again the loss of expressive power between quantitative and qualitative representations for the values of quantities. A quantitative representation implies an infinite number of possible values for quantities. Under such a representation there can be no ambiguous experiments because of the pigeonhole principle. However, there is a tradeoff: the size of the search space of experiments grows as a power of the number of possible predictions.

### 6.1.2 Experiments Based on Transitions

Experiments can be designed from knowledge of how *transitions* between the values of quantities [2] combine, as well as from knowledge of how the values themselves combine. Transitions are specified by a previous value, a new value, and a direction of change of a quantity. Predictions of transitions are constrained only by the current value of a quantity. In general, there are a maximum of five possible transitions for a quantity.

The design of an experiment based on transitions for the same example as above proceeds as follows: (Assume that the current value of $A$ is 0; of $B$ is +; of $C$ is −).

| SITUATION | | | PREDICTION OF | PREDICTION OF |
|---|---|---|---|---|
| $A$ | $B$ | $C$ | $(-A) \times (B + C)$ | $A \times (B \times C)$ |
| ST0 | ST+ | ST- | ST0 | ST0 |
| ST0 | ST+ | AD- | ST0 | ST0 |
| ST0 | ST+ | SB- | ST0 | ST0 |
| . | | | | |
| . | (as long as $A$ is ST0, all predictions are ST0) | | | |
| . | | | | |
| EN+ | ST+ | ST- | ST0 EN+ EN- | EN- |
| EN+ | ST+ | AD- | ST0 EN+ EN- | EN- |
| EN+ | ST+ | SB- | ST0 EN+ EN- | EN- |
| EN+ | ST+ | DB- | EN- | ST0 |

After a string of hopelessly ambiguous experiments, three are found that are only potentially ambiguous and finally one that is unambiguous.

As is the case for experiments based on the values of quantities, the search for an experiment based on transitions of quantities can be heuristically constrained by utilizing higher-level knowledge of constraints on the ways in which transitions combine under negation, mulitplication, and addition. For example, knowledge that the result of multiplication by ST0 is always ST0 can prevent the experiment designer from considering many situations which must give rise to hopelessly ambiguous experiments.

Transitions offer a tradeoff concerning ambiguity. Although the ambiguity due to the pigeonhole principle is slightly ameliorated because the number of possible predictions increases from three to five, the ambiguity due to indeterminism in the calculi is greater, as can be easily seen from an inspection of the transition calculi in the appendix.

### 6.1.3 Experiment Design via Constraint Back-Propagation

Experiment design also can be approached by constraining predictions rather than situations. In other words, another way to design an experiment is to enumerate assignments of predictions to hypotheses and determine if there is a single situation which implies all those predictions.

Using the same example as above, we can enumerate the assignments of value predictions [3] to the two hypotheses $E = (-A) \times (B + C)$ and $E = A \times (B \times C)$ which correspond to unambiguous experiments.

[2] e.g. STAY+, DISABLE-, etc. See the appendix.

[3] This alternate approach to experiment design using constraint back-propagation applies equally well to predictions of *transitions* between values of quantities.

| PREDICTION OF $(-A) \times (B + C)$ | PREDICTION OF $A \times (B \times C)$ |
|---|---|
| 0 | + |
| 0 | − |
| + | 0 |
| + | − |
| − | 0 |
| − | + |

This alternate approach to experiment design involves searching through a space of prediction assignments for hypotheses which is exponential in the number of hypotheses to be distinguished. [4] The enumeration of prediction assignments can halt at an assignment for which a situation can be found which predicts all the desired values of the assignment. This situation is the design for an unambiguous experiment.

The situations which manifest a particular desired value for a particular hypothesis can be found by *constraint back-propagation*. The predicted value for the dependent quantity is back-propagated through the constraints corresponding to the composition of operators in the hypothesis to determine the assignments of values to the independent quantities which result in the desired value for the dependent quantity.

Consider the assignment of predictions in the fourth line of the table above. We must determine the assignments of values to $A$, $B$, and $C$ which result in a value of + for $E$ in the hypothesis $E = (-A) \times (B + C)$ and in a value of − for $E$ in the hypothesis $E = A \times (B \times C)$. Any common assignment is the design for an unambiguous experiment.

The constraint back-propagation proceeds as follows:

The first relevant constraint is "A product is + when both multiplicands are the same and neither is 0."

| $-A$ | $B + C$ | $(-A) \times (B + C)$ |
|---|---|---|
| + | + | + |
| − | − | |

The next relevant constraint is the definition of negation.

| $A$ | $-A$ |
|---|---|
| − | + |
| + | − |

The next constraint is "A sum is positive/negative when at least one operand is positive/negative and neither is negative/positive."

| $B$ | $C$ | $(B + C)$ |
|---|---|---|
| 0 | + | + |
| + | 0 | |
| + | + | |
| 0 | − | − |
| − | 0 | |
| − | − | |

[4] It had better be exponential in something; otherwise this reformulation of the experiment design problem would be suspicious.

These constraints are combined via back-propagation.

| $A$ | $B$ | $C$ | $(-A) \times (B + C)$ |
|---|---|---|---|
| − | 0 | + | + |
| − | + | 0 | |
| − | + | + | |
| + | 0 | − | |
| + | − | 0 | |
| + | − | − | |

Constraint back-propagation also is applied to the other hypothesis.

| $A$ | $B$ | $C$ | $A \times (B \times C)$ |
|---|---|---|---|
| + | + | − | − |
| + | − | + | |
| − | + | + | |
| − | − | − | |

All that remains to be done is to determine if there is an assignment which predicts the desired values for both hypotheses. There is exactly one, namely $(A/-)(B/+)(C/+)$. This is the sought-after design for an unambiguous experiment. If the result of this experiment (the value of $E$) is +, then $(-A) \times (B + C)$ is the correct hypothesis; if the result is −, then $A \times (B \times C)$ is the correct hypothesis.

### 6.1.4 Ambiguity Again

Ambiguity still must be dealt with in this alternate approach to experiment design. Ambiguity due to indeterminism in the calculi also appears during constraint-back propagation. For example, the constraint concerning positive sums stated above as "A sum is positive/negative when at least one operand is positive/negative and neither is negative/positive" has a more general form which reads "A sum may be positive/negative when at least one operand is positive/negative." This more general constraint reflects the fact that the assignment of one positive operand and one negative operand may also produce negative or zero sums.

Ambiguity due to indeterminism in the calculi can be removed during constraint back-propagation by simply pruning out assignments which have indeterminate predictions. This was done in the example above by using the constraint concerning positive sums which implicitly prunes out ambiguous assignments.

On the other hand, there is nothing to be done about ambiguity due to the pigeonhole principle. This ambiguity appears before constraint back-propagation, during the assignment of predictions to hypotheses. There simply is no way to assign $p$ predictions to $h$ hypotheses, $p > h$, such that no prediction is repeated.

## 6.2 Experiments Based on Knowledge of Causal Mechanisms

The mechanistic hypotheses which are derivable from the domain theory may be only partially justified because of uninstantiable assertions or because of intrinsic approximations in the domain

theory. For this reason, the learning system may propose more than one partially justified hypothesis to explain the same observation. The same constraint back-propagation technique used to design experiments to distinguish entirely empirical hypotheses also can be used to design experiments to distinguish partially justified, mechanistic hypotheses.

For example, suppose that the causal modelling system is presented with the following observation

$$\begin{aligned}
&\ldots QuantityOf(Taillight1, Intensity1)\\
&QuantityOf(Sun, Intensity2)\\
&PhysicalConnection(Automobile1, Taillight1)\\
&ValueOf(Intensity1, Bright)\\
&ValueOf(Intensity2, VeryBright) \ldots
\end{aligned}$$

and the causal reasoning task of explaining (finding the cause of) the glowing taillight.

The causal modelling system can propose two different causal explanations for the brightness of an object, one involving light transmission and the other involving the electrophotic transformation and electricity. Both of these causal explanations can be partially instantiated into the above observation.

An explanation in terms of light transmission looks like:

$$\begin{aligned}
&\exists(m)\neg\exists(b)\\
&\quad (Intensity(Intensity2) \wedge Intensity(Intensity1) \wedge\\
&\quad StraightLinePath(m) \wedge Between(m, Sun, Taillight1) \wedge\\
&\quad Opaque(b) \wedge Along(m, b))\\
&\Longrightarrow FunctionalDependence(Intensity2, Intensity1)
\end{aligned}$$

An explanation in terms of the electrophotic transformation and electricity looks like: [5]

$$\begin{aligned}
&\exists(q1)\\
&\quad (ElectricCharge(q1) \wedge Intensity(Intensity1))\\
&\Longrightarrow FunctionalDependence(q1, Intensity1)
\end{aligned}$$

$$\begin{aligned}
&\exists(q2)\exists(q1)\neg\exists(b)\\
&\quad (ElectricCharge(q2) \wedge ElectricCharge(q1) \wedge\\
&\quad ElectricalConductor(Wire1) \wedge Between(Wire1, Battery1, Taillight1) \wedge\\
&\quad (Discontinuity(b) \vee ElectricalInsulator(b)) \wedge Along(Wire1, b))\\
&\Longrightarrow FunctionalDependence(q2, q1)
\end{aligned}$$

Both of these explanations are partially instantiated, and hence partially justified. The task now is to design an experiment which can gather further evidence to distinguish them.

Experiment design proceeds in several steps. First, different predictions are assigned to the two hypotheses. Then constraint back-propagation determines the instantiations for the two hypotheses under which their assigned predictions can be realized. Finally, an instantiation common to both hypotheses is chosen as the experiment.

Suppose that the light transmission hypothesis is assigned a prediction stating that the value of the intensity of the taillight is 0. Further suppose that the electrophotic transformation hypothesis is assigned a prediction stating that the value of the intensity of the taillight is +.

[5] The instantiation of this explanation requires perceptions inside the automobile.

Using constraint back-propagation, the range of situations for light transmission in which the intensity of the taillight turns out to be 0 is determined to be:

$$= (Intensity2, 0) \vee$$

$$\neg\exists(m)$$
$$(= (Intensity2, +) \wedge$$
$$StraightLinePath(m) \wedge Between(m, Sun, Taillight1)) \vee$$

$$\exists(m)\exists(b)$$
$$(= (Intensity2, +) \wedge$$
$$StraightLinePath(m) \wedge Between(m, Sun, Taillight1) \wedge$$
$$Opaque(b) \wedge Along(m, b))$$

These are the situations in which the intensity of the sun is zero or the light transmission mechanism is inactive.

The range of situations for the electrophotic transformation in which the intensity of the taillight turns out to be *positive* is determined to be:

$$\exists(q2)\exists(q1)\neg\exists(b)$$
$$(ElectricCharge(q2) \wedge = (q2, +) \wedge$$
$$ElectricCharge(q1) \wedge$$
$$ElectricalConductor(Wire1) \wedge Between(Wire1, Battery1, Taillight1) \wedge$$
$$(Discontinuity(b) \vee ElectricalInsulator(b)) \wedge Along(Wire1, b))$$

These are the situations in which the battery does generate a current and the electricity and electrophotic transformation mechanisms are active.

Finally, the sought-after experiment which can further distinguish these two hypotheses is any instantiation which satisfies the back-propagated constraints for both hypotheses, such as:

$$\exists(q2)\exists(q1)\neg\exists(b)\exists(m)$$
$$(ElectricCharge(q2) \wedge = (q2, +) \wedge$$
$$ElectricCharge(q1) \wedge$$
$$ElectricalConductor(Wire1) \wedge Between(Wire1, Battery1, Taillight1) \wedge$$
$$(Discontinuity(b) \vee ElectricalInsulator(b)) \wedge Along(Wire1, b) \wedge$$
$$= (Intensity2, +) \wedge$$
$$StraightLinePath(m) \wedge Between(m, Sun, Taillight1) \wedge$$
$$Opaque(Card1) \wedge Along(m, Card1))$$

If the intensity of the taillight turns out to be zero when this experiment is executed, then the light transmission hypothesis is supported; if the intensity of the taillight turns out to be positive, then the electrophotic transformation hypothesis is supported.

### 6.2.1 Uninstantiable and Uncertain Experiments

The results of experiments derived from the domain theory ultimately suffer from the same lack of certainty as do hypotheses generated from the domain theory. The correct interpretation of the

results of an experiment may depend on an uninstantiable assertion or on an exceptional case which is hidden by an approximation. Experiments do actively gather greater justification for hypotheses, but they can never completely erase their empirical nature.

In the example above, a disrupting magnetic field may be the exceptional explanation for an outcome in which the intensity of the taillight is zero. This possible explanation does not appear in the abstracted electricity schema. Another possible interpretation of the intensity of the taillight being zero is that the battery is not generating any current. An assertion that the battery is generating current does appear in the specification of the experiment, but the current does not appear in a perception, and this assertion remains uninstantiated.

An assertion in the specification of an experiment may be uninstantiable for several reasons: it is unobserved (e.g. the current); it is intrinsically unobservable (e.g. the heat path in a toaster); or it is *unachieveable*, i.e., it has an unalterable instantiation other than the desired one. The experiment specification above suggests setting the intensity of the sun to zero as one way to render the light transmission mechanism inactive. This is not an achieveable assertion. On the other hand, an opaque barrier placed between the sun and the taillight is an achieveable assertion.

A planner can determine (or fail to determine) how to achieve assertions specified in the design of an experiment. The issues involved in the interaction of an experiment designer and a planner deserve further exploration.

### 6.2.2 Experiments for a Single Hypothesis

Experiments designed to distinguish hypotheses gather more justification for the hypothesis(es) whose predictions are corroborated. The justification amplification associated with experimenting also can be applied to a single hypothesis. This use of experimenting is particularly appropriate when there is a single relevant *known* causal mechanism which can explain an observation, but which could not be fully instantiated.

The way to further justify a single hypothesis is to design an experiment which is more fully instantiable (possibly at a more detailed level of explanation) than the current version of the hypothesis. The experiment may generate a prediction which is different from previous observations. If the prediction is corroborated, the hypothesis is now more strongly justified.

For example, suppose that the scenario above involves a taillight sitting in the middle of the street, unattached to any automobile, so that the only relevant hypothesis is the one involving light transmission. The experiment involving the placement of an opaque card between the sun and the taillight can make this single hypothesis more justified. An *instantiated* light transmission barrier is now part of the hypothesis; furthermore, it is now possible to explain why the taillight sometimes is dark.

## 6.3 Summary

This chapter can be summarized in the following observations towards a set of principles for learning from experiments.

- Experiment design can proceed either by searching for a situation for which there is a divergent set of predictions, or by searching for a set of divergent predictions for which there is a common situation.

- Experiments can be designed either for hypotheses with a purely empirical basis, or for hypotheses partially justified by explanations derived from a domain theory.
- Experiments can distinguish several hypotheses or can more strongly justify a single hypothesis.
- Most experiments are ambiguous – it is difficult to design an experiment in which every competing hypothesis predicts a different result.
- The best experiments are those that involve only instantiable (observable and achieveable) assertions.
- The interpretation of the results of an experiment can be uncertain because the experiment is not fully instantiated or because the experiment is derived from an approximate domain theory.

# Chapter 7

# Constructing, Using, and Refining Causal Models

In this section, I present the procedure for constructing causal explanations/descriptions/models to support a causal reasoning task. This procedure makes use of the representations for causality, the inductive and deductive inference methods, and the domain theory presented in previous sections. I demonstrate also how the causal models produced by this procedure support their motivating causal reasoning tasks. Finally, I show how causal models which prove to be inconsistent may be repaired.

## 7.1 Specifying the Causal Reasoning Task

Some examples of causal reasoning tasks in the domain of physical systems are:

- (Planning involving a camera). How to control film exposure by using the aperture ring, a flash, and/or different film?
- (Prediction involving a bathtub). What will happen if the drain is plugged and the faucet turned?
- (Diagnosis involving a toaster). Why did the bread not turn to toast?

The most conceptually simple causal reasoning tasks involve searches through causal networks, beginning from certain nodes specified *a priori* to be causes or effects. These search problems can be stated as one of the following questions:

- What are the effects of these causes?
- What are the causes of these effects?
- What are the causal relations between these causes and effects?

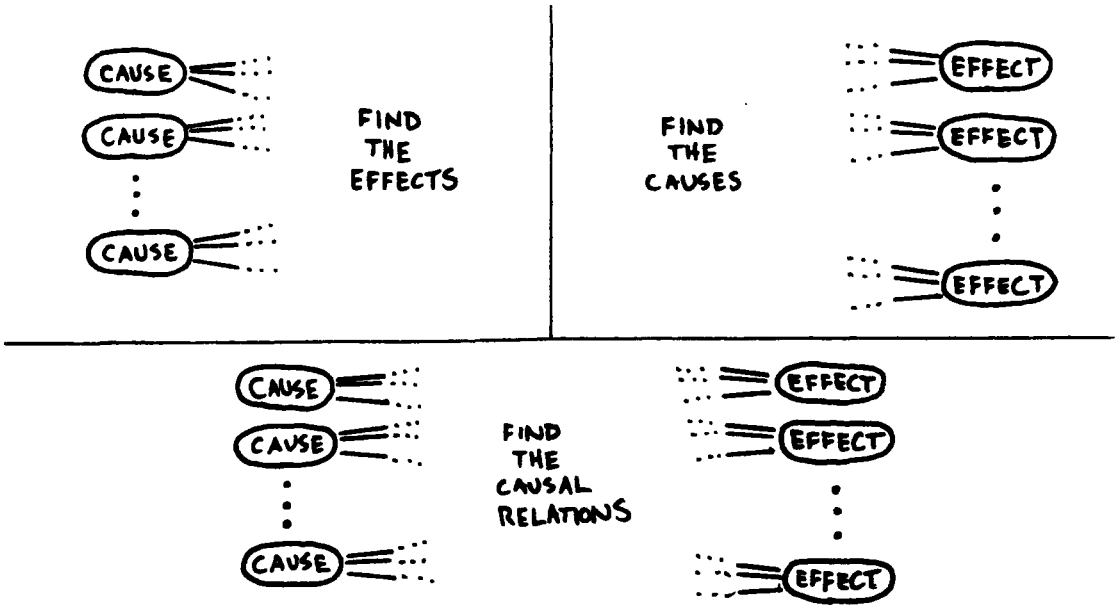


Figure 7.1: How the Causal Reasoning Task Constrains Causal Modelling

The planning problem above involves finding causal relations between the actions of changing the setting of the aperture ring, changing the intensity of a flash bulb, and changing the type of film (all taken to be causes), and the exposure on a piece of film (taken to be an effect). The prediction and diagnosis problems above can be similarly interpreted as a search problem in a causal network.

All of the causal reasoning tasks which provide contexts for my causal modelling system are presented in this form. The task of the causal modelling system is to produce causal networks (models) on which the motivating causal reasoning tasks can be performed successfully. This is the yardstick of success for the causal modelling process.

The motivating causal reasoning tasks serve to constrain the causal modelling process by providing anchor points (the specified causes and/or effects) around which the construction of causal models can begin. The modelling process can be terminated heuristically, by computational resource constraints, or by exhausting the observations. When both causes and effects are specified, causal modelling is terminated when all specified causes and effects have been causally linked. There is a degenerate case where no causes and effects are specified. In this case, an unconstrained causal modelling task is generated where the initial causes are taken to be the earliest events in the observations and the final effects are taken to be the latest events.

## 7.2 The Causal Modelling Procedure

The following is the procedure for constructing causal models in the context of a causal reasoning task. Step 6 can be executed in the background. Step 8 is non-deterministic.

Given a causal reasoning task concerning a physical system: (The causal reasoning task is specified as a set of causes and/or effects).

1. Procure an observation of the physical system.

An observation is a structural and behavioral description of the physical system over time.

2. Generate primitive causal hypotheses relevant to the causal reasoning task.

These are proportionality relations (the simplest functional dependencies) between quantities specified as causes and/or effects in the causal reasoning task, or between quantities already shown to be causally related to a cause and/or effect specified in the causal reasoning task.

3. Remove those hypotheses which do not satisfy causal direction constraints.

Effects must be after or simultaneous with causes. Known actions can be only causes. Primitive causes and effects specified in the causal reasoning task cannot be other than as specified.

4. Retrieve and instantiate causal mechanism schemata.

For each primitive causal hypothesis, retrieve relevant causal mechanism schemata. Schemata are indexed under the types of cause and effect pairs they can explain.

For each retrieved causal mechanism schema, attempt to instantiate the schema to a level which generates a mechanistic, justified hypothesis/explanation which is consistent with the observation. A schema may not be fully instantiable because of unobserved assertions.

5. Generate compositions of causal hypotheses.

Compositions of causal hypotheses can be generated empirically by combining functional dependencies under negation, multiplication, and addition and corroborating the expected values of quantities. Justified compositions of hypotheses are generated when schemata already containing compositions are instantiated or when causal mechanism schema instantiations intersect.

6. Learn new compositions of causal mechanisms.

If there are instantiated causal mechanism schema which intersect, then generalize the constraint(s) at the intersections and index the new composed causal mechanisms under the types of causes and effects they explain.

7. Is the causal model adequate?

If there remain competing hypotheses, then go to 8.

If there are causes and/or effects specified in the causal reasoning task which have not yet been incorporated into the causal model, and other termination conditions have not been reached, then go to 2.

Otherwise, stop.

8. Open up the system.

If there are causal hypotheses for which there are relevant causal mechanisms which could not be instantiated, then procure an observation at a finer level of perceptual granularity. Go to 2.

8. Design experiments to distinguish hypotheses.

If there are competing causal hypotheses, purely empirical or partially justified, then design an experiment which can distinguish the hypotheses. Designing an experiment involves determining a situation for which the competing hypotheses generate divergent predictions. Instantiate the situation, procure an observation and eliminate hypotheses whose predictions are not corroborated. Go to 7.

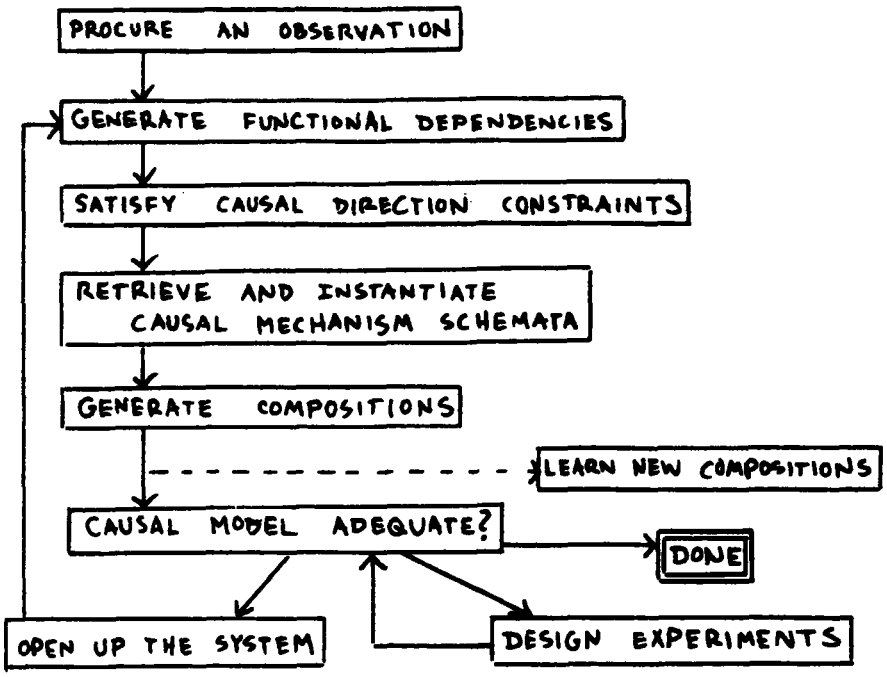


Figure 7.2: The Causal Modelling Procedure

8. Design experiments to further justify hypotheses.

If there are causal hypotheses for which there are relevant but uninstantiable causal mechanisms which involve inherently unobservable assertions, then determine situations which involve observable assertions which can further justify the causal hypotheses. Instantiate the situations, procure observations and further justify those hypotheses whose predictions are corroborated. Go to 7.

This procedure is depicted in flowchart fashion in Figure 7.2.

## 7.3 Constructing a Causal Model

In this section, I present a somewhat detailed example of how the learning system, using the above procedure, constructs a causal model of a camera.

The causal reasoning task which provides context, constraint, and motivation for the causal modelling system is specified below:

Determine causal relations between these causes:

The f-stop setting of the aperture ring.
The speed of the film.

and this effect:

The exposure on the film.

1. Procure an observation.

The following is a structural and behavioral description of the camera over time. This observation describes the physical objects and quantities which make up the camera and how they change.

.
.
.

$QuantityOf1(FStop1, ApertureRing1)$
$QuantityOf2(Intensity1, Subject1)$
$QuantityOf3(Length1, Lens1)$
$QuantityOf4(Speed1, Film1)$
$QuantityOf5(Exposure1, Film1)$
$QuantityOf6(Distance1, FocusRing1)$
$QuantityOf7(Position1, Release1)$
$QuantityOf8(Intensity2, Flash1)$
$QuantityOf9(Location1, Subject1)$

$PhysicalConnection1(ApertureRing1, Lens1)$
$PhysicalConnection2(FocusRing1, Lens1)$
$PhysicalConnection3(Lens1, Camera1)$
$PhysicalConnection4(Flash1, Camera1)$
$Inside1(Film1, Camera1)$
$PhysicalConnection5(Release1, Camera1)$

$ValueOf1(FStop1, t1, 5.6)$
$ValueOf2(Intensity1, t1, Dim)$
$ValueOf3(Length1, t1, Long)$
$ValueOf4(Speed1, t1, 400)$
$ValueOf5(Exposure1, t1, Blank)$
$ValueOf6(Distance1, t1, 3)$
$ValueOf7(Position1, t1, Up)$
$ValueOf8(Intensity2, t1, Dark)$
$ValueOf9(Location1, t1, (0, 20, 3))$

$ValueOf3(Length1, t2, Short)$
$ValueOf6(Distance1, t2, 20)$

$ValueOf2(Intensity1, t3, Bright)$
$ValueOf7(Position1, t3, Down)$
$ValueOf8(Intensity2, t3, Bright)$

$ValueOf7(Position1, t4, Up)$

$ValueOf5(Exposure1, t5, OverExposed)$

.
.
.

2. Generate primitive causal hypotheses relevant to the specified causal reasoning task.

The causes and effects specified in the causal reasoning task form the initial focus for causal modelling. $FStop1$ and $Speed1$ are specified as causes; $Exposure1$ is specified as an effect.

3. Remove those hypotheses which do not satisfy causal direction constraints.

$FStop1$ and $Speed1$ can be only causes. $Exposure1$ can be only an effect. $Distance1$ and $Position1$ represent external controls of the camera; these can be only causes. Finally, effects cannot precede causes in time.

The generated primitive causal hypotheses are:

$FunctionalDependence(FStop1, Intensity1)$
$FunctionalDependence(FStop1, Length1)$
$FunctionalDependence(FStop1, Exposure1)$
$FunctionalDependence(FStop1, Intensity2)$
$FunctionalDependence(Speed1, Intensity1)$
$FunctionalDependence(Speed1, Length1)$
$FunctionalDependence(Speed1, Exposure1)$
$FunctionalDependence(Speed1, Intensity2)$
$FunctionalDependence(Intensity1, Exposure1)$
$FunctionalDependence(Length1, Exposure1)$
$FunctionalDependence(Distance1, Exposure1)$
$FunctionalDependence(Position1, Exposure1)$
$FunctionalDependence(Intensity2, Exposure1)$

4. Justify hypotheses by instantiating causal mechanism schemata.

None of these hypotheses can be justified as yet; some because they are incorrect, others because they require observations of objects and quantities not available at the current level of perceptual resolution.

5. Generate compositions of causal hypotheses.

Only empirical compositions at the level of combined functional dependencies can be generated because no causal mechanism schemata have been instantiated. Among those generated are:

.
.
.
$FunctionalDependence(FStop1 + Speed1, Intensity1)$
$FunctionalDependence(FStop1 \times Intensity1 \times Speed1, Exposure1)$
$FunctionalDependence(FStop1 \times Intensity1, Exposure1)$
$FunctionalDependence((Speed1 + Intensity2) \times Length1, Exposure1)$
.
.
.

At this point, the causal modelling system is suffering from its inability to instantiate any causal mechanism schemata and focus search on the resulting justified hypotheses. Without this focussing, the empirically generated hypothesis space is exponentially large. [1]

6. Learn new compositions of causal mechanisms.

No causal mechanism schemata have been instantiated.

7. Is the causal model adequate?

There are many competing hypotheses, so ...

8. Open up the system or design experiments.

The causal modelling system now can design experiments to distinguish the primitive and composed causal hypotheses or it can procure an observation at a finer level of perceptual granularity and once again try to instantiate causal mechanism schemata. "Knob-tweaking" experiments would be able to gather empirical evidence that, for example, the film exposure depends on the f-stop, but not on the setting of the focus ring. However, given that there are currently an exponential number of hypotheses, another attempt to instantiate some causal mechanism schemata, justify some causal hypotheses, and focus search, is indicated.

Upon opening up the camera, the causal modelling system does indeed procure perceptions of physical objects and quantities which were unavailable before. Among these new perceptions are:

.
.
.

$QuantityOf10(Intensity3, Film1)$
$QuantityOf11(Area1, Iris1)$
$QuantityOf12(Position2, Shutter1)$
$QuantityOf13(Location2, Film1)$
$QuantityOf14(Location3, Iris1)$
$QuantityOf15(Location4, Shutter1)$

$Inside2(Iris1, Lens1)$
$Inside3(Shutter1, Camera1)$
$PhysicalConnection6(ApertureRing1, Iris1)$
$PhysicalConnection7(Release1, Shutter1)$

$ValueOf10(Intensity3, t1, Dark)$
$ValueOf11(Area1, t1, 4)$
$ValueOf12(Position2, t1, In)$
$ValueOf13(Location2, t1, (0, 0, 3))$
$ValueOf14(Location3, t1, (0, 0.5, 3))$
$ValueOf15(Location4, t1, (0, 0.1, 3))$

$ValueOf10(Intensity3, t3, Bright)$
$ValueOf12(Position2, t3, Out)$
$ValueOf15(Location4, t3, (0.2, 0.1, 3))$

[1] Exponential in the number of quantities appearing in observations.

$ValueOf12(Position2, t4, In)$
$ValueOf15(Location4, t4, (0, 0.1, 3))$
.
.
.

2. Generate primitive causal hypotheses relevant to the specified causal reasoning task.

3. Remove those hypotheses which do not satisfy causal direction constraints.

Among the new primitive causal hypotheses are:

.
.
.
$FunctionalDependence(Intensity1, Intensity3)$
$FunctionalDependence(Area1, Intensity3)$
$FunctionalDependence(Position2, Intensity3)$
$FunctionalDependence(FStop1, Area1)$
$FunctionalDependence(Position1, Position2)$
$FunctionalDependence(Intensity3, Exposure1)$
.
.
.

4. Justify hypotheses by instantiating causal mechanism schemata.

This time around, several causal mechanism schemata can be retrieved and instantiated:

The proposed functional dependence between the intensity and the exposure of the film can be explained by a photochemical transformation on the film.

$$\begin{aligned} & Intensity(Intensity3) \wedge Appearance(Exposure1) \wedge = (Film1, Film1) \\ \Longrightarrow\ & FunctionalDependence(Intensity3, Exposure1) \end{aligned}$$

The light transmission schema can be retrieved by the proposed functional dependence between the intensity of the subject and the intensity of the film, or by the proposed functional dependence between the area of the iris and the intensity of the film, or by the proposed functional dependence between the position of the shutter and the intensity of the film.

$$\begin{aligned} & Intensity(Intensity1) \wedge Intensity(Intensity3) \wedge \\ & StraightLinePath(Path1) \wedge Between(Path1, Subject1, Film1) \wedge \\ & Opaque(Iris1) \wedge Along(Iris1, Path1) \wedge \\ & Opaque(Shutter1) \wedge Along(Shutter1, Path1) \\ \Longrightarrow\ & FunctionalDependence(Intensity1 \times Area1 \times Position2, Intensity3) \end{aligned}$$

Knowledge of geometry is needed to support some of the inferences which contribute to the instantiation of this schema. *Path*1 and the *Along* assertions can be computed from the location quantities of the subject, iris, shutter, and film. There is a straight line through all of these points. The cross-sectional areas of the light path passed by the iris and the shutter can be computed from their location quantities using projection methods.

I have not decided if my domain theory will include this "strong" knowledge of geometry. Without it, these assertions would be uninstantiable. For the purposes of this example, I assume the causal modelling system does have access to this knowledge.

The proposed functional dependence between the f-stop and the area of the iris can be explained by a mechanical coupling:

$$\begin{aligned} & Position(FStop1) \wedge Position(Area1) \wedge \\ & PhysicalConnection(PhysicalConnection6) \wedge \\ & Between(PhysicalConnection6, ApertureRing1, Iris1) \\ \Longrightarrow\ & FunctionalDependence(FStop1, Area1) \end{aligned}$$

Finally, the proposed functional dependence between the positions of the release and shutter also can be explained by a mechanical coupling.

$$\begin{aligned} & Position(Position1) \wedge Position(Position2) \wedge \\ & PhysicalConnection(PhysicalConnection7) \wedge \\ & Between(PhysicalConnection7, Release1, Shutter1) \\ \Longrightarrow\ & FunctionalDependence(Position1, Position2) \end{aligned}$$

The domain theory contains no schema or composition of schemata which can explain how the speed of the film might affect the exposure of the film. [2]

5. Generate compositions of causal hypotheses.

Several justified compositions of causal hypotheses have been generated for free. The instantiated light transmission schema contains a composition of three functional dependencies, all sharing the intensity of the film as dependent quantity. In addition, the light transmission schema intersects with the photochemical transformation schema at the intensity of the film; with one mechanical coupling at the area of the iris; with the other mechanical coupling at the position of the shutter.

Because there are no known causal mechanisms in which the film speed might be participating, the causal modelling system can generate compositions involving the film speed only empirically. However, any proposed contribution of the film speed should now be constrained to intersect with the justified causal model of the camera already constructed. Among the possible composition hypotheses involving the film speed are: [3]

[2] The correct explanation is: The speed or ASA rating of a roll of film reflects the density of its emulsion layer. The denser this layer, the slower the photochemical reaction. This is an integration argument.

[3] Compositions containing causal cycles are not generated.

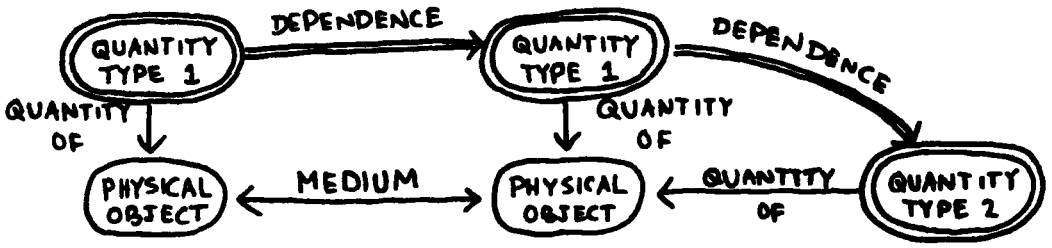


Figure 7.3: A Transformation Across Two Objects

$$
\begin{aligned}
&FunctionalDependence \\
&\quad (Speed1 + (Intensity1 \times Area1 \times Position2), Intensity3) \\
&FunctionalDependence \\
&\quad (Speed1 \times (Intensity1 \times Area1 \times Position2), Intensity3) \\
&FunctionalDependence(Speed1 + Intensity3, Exposure1) \\
&FunctionalDependence(Speed1 \times Intensity3, Exposure1)
\end{aligned}
$$

6. Learn new compositions of causal mechanisms.

The intersecting causal mechanism schema instantiations in the causal model of the camera can be lifted out, generalized, and added to the domain theory under appropriate indexes. For example, the intersection of the light transmission and photochemical transformation schemata can be generalized to a transformation across two objects, involving a propagation across two objects and a transformation at the second object. The constraint which must hold is that a single quantity of the second object must serve as both dependent quantity of the propagation and independent quantity of the transformation.

Both of the intersections between a mechanical coupling and the light transmission can be generalized to a composed *valve* mechanism where one object serves as both the dependent half of a mechanical coupling and as barrier to a flow. (See Figure 3.8).

7. Is the causal model adequate?

There are still competing hypotheses concerning the contribution of the film speed on the film exposure.

8. Open up the system or design experiments.

The film is a primitive physical object which cannot be taken apart; thus opening up the system is not an option. The final task of the causal modelling system is to design experiments to distinguish the hypotheses concerning the film speed. Recall these hypotheses:

$$
\begin{aligned}
&FunctionalDependence \\
&\quad (Speed1 + (Intensity1 \times Area1 \times Position2), Intensity3) \\
&FunctionalDependence \\
&\quad (Speed1 \times (Intensity1 \times Area1 \times Position2), Intensity3) \\
&FunctionalDependence(Speed1 + Intensity3, Exposure1) \\
&FunctionalDependence(Speed1 \times Intensity3, Exposure1)
\end{aligned}
$$

Using knowledge that the results of addition and multiplication are different when exactly one operand is zero, the causal modelling system designs the following experiment:

The hypotheses where the film speed makes an additive contribution are assigned the prediction $+$. The hypotheses where the film speed makes a multiplicative contribution are assigned the prediction 0. Constraint back-propagation determines the situations which predict the assigned outcomes for all hypotheses: Either the intensity of the subject must be zero, or the iris must be closed completely, or the shutter must not be moved from in front of the film. The causal modelling system does not know how to control the intensity of the subject, nor can the iris be closed completely; these experiment designs are not operational. However, the shutter can be kept in front of the film by not pushing the release button. If the film speed makes an additive contribution, then the exposure should not come out blank (zero), even though no light falls on the film.

Of course, the film does come out blank. Now there remain only these hypotheses:

$$
\begin{aligned}
&FunctionalDependence \\
&\quad (Speed1 \times (Intensity1 \times Area1 \times Position2), Intensity3) \\
&FunctionalDependence(Speed1 \times Intensity3, Exposure1)
\end{aligned}
$$

The intensity of the film is affected by the film speed in only one of these hypotheses. Thus a simple experiment is to change the film speed and determine if the film intensity also changes. Under an assigned transition of ADD+ for the film speed, the first hypothesis predicts an ADD+ transition for the film intensity while the second hypothesis predicts STAY+.

The result of the experiment is STAY+ and the first hypothesis is eliminated.

7. Is the causal model adequate?

There is now a single justified causal model of the camera. All of the causes and effects in the motivating causal reasoning task (the f-stop, the film speed, and the film exposure) have been incorporated into this model. The causal model is adequate.

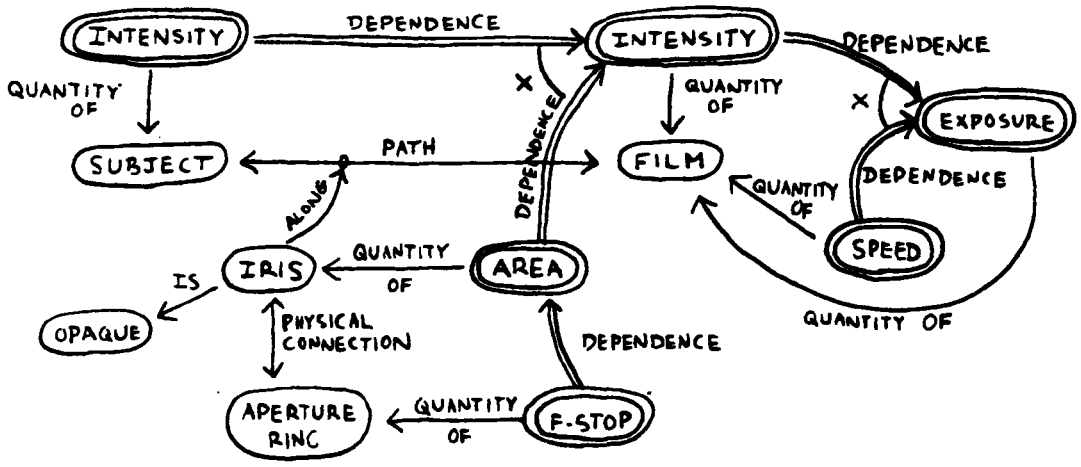


Figure 7.4: The Causal Model of the Camera

## 7.4 Using and Refining a Causal Model

Causal modelling takes place in the context of a causal reasoning task. The criterion of success for the causal modelling system is: The causal model delivered by the learning system must be able to support the motivating causal reasoning task.

This is easy to demonstrate for the case of the camera. The motivating causal reasoning task is a planning problem: How to control film exposure by controlling the f-stop and/or the film speed. In particular, suppose the film comes out overexposed. The planning problem now becomes: How to decrease the film exposure by controlling the f-stop and/or the film speed.

The same constraint back-propagation method used in experiment design can double as a simple planner. The film exposure is assigned a transition of SUBTRACT+, and this constraint is back-propagated through the network of causal relations in the causal model. The result is a set of transition assignments to quantities which are causally "upstream" from the film exposure. Those assignments which specify changes in the f-stop and/or the film speed are the desired plans.

The causal models generated by the causal modelling system may prove to be inconsistent because of incomplete instantiations or because of approximations in the domain theory. In some cases, it is possible to refine an inconsistent causal model by making another attempt at instantiation at a more detailed level of explanation.

For example, suppose that a lens cap is inadvertently placed on the camera while a photograph is taken. The expected outcome that the film has some positive exposure is not corroborated. However, the means of repairing the model of the camera already exists in the light transmission schema. The lens cap can be instantiated as yet another opaque physical object along the path from the subject to the film acting as a barrier to light. Now the model generates the correct prediction that no light reaches the film and the film comes out blank.

In the summary chapter of this report, I make further arguments that my causal models can support causal reasoning tasks. In particular, I argue that they can support qualitative reasoning (e.g. envisioning, limit analysis) and fault diagnosis. It is not my purpose to build state-of-the-art versions of these types of causal reasoning systems, but I must argue that my causal modelling system can interface with them in the intended manner – as a front end producing the representations on which they operate.

# Chapter 8

# Summary

In this work, I have shown how the causal modelling competence can arise from a combination of inductive and deductive inference, employing knowledge of the general form of causal relations and of the kinds of causal mechanisms that exist in a domain. The hypotheses generated by such a learning system range from purely empirical to more and more strongly justified.

In this final chapter, I discuss how my work relates to other research efforts and I recount the proposed solutions to the set of issues which my research addresses.

## 8.1 Relation to Other Work

In this section, I discuss how my research relates to the research of others who have worked in the following areas: qualitative representations for and qualitative reasoning about physical systems, including fault diagnosis; empirical and analytical learning; levels of abstraction; learning from experiments.

### 8.1.1 Qualitative Reasoning about Physical Systems

My work on causal modelling is intended to interface in a direct way with work on qualitative reasoning about physical systems. Qualitative reasoning systems are given process descriptions [Forbus 84] or component and device models [de Kleer & Brown 84]. Using these descriptions, they perform certain reasoning tasks such as *envisioning*, i.e., predicting the possible next state(s) of physical systems. My causal modelling system is intended to acquire the descriptions on which these qualitative reasoning systems operate. In the next few paragraphs, I argue that my causal models are the right kind of descriptions which can support qualitative reasoning tasks.

There is a strong resemblance between Forbus' process descriptions and my causal mechanism descriptions, from which my causal models are derived. This is certainly the case at the level of quantities and functional dependencies because I have adopted Forbus' representations directly. At the level of physical structure, I believe my representation is more perspicuous. I distinguish between media and barriers while Forbus subsumes these terms into a generic category of *physical*

*precondition*. The identification of media and barriers is directly relevant to the task to determining which processes (or causal mechanisms) are, repsectively, active or inactive.

Forbus' representations support *limit analysis* - determining what process changes occur because of changes in the qualitative values of quantities. Determining the next value of a quantity is supported by a simple notion of time integrability which states that the next value of a quantity is the adjacent value in the space of values for the quantity in the direction indicated by the rate of the quantity.

Before time integration and limit analysis can be applied to my causal models, the interesting *limit points*, or *thresholds* (given *a priori* in Forbus' process descriptions) have to be identified. Some of these thresholds may come for free (e.g., zero points); I claim that others can be identified by recognizing media and barriers. Processes change exactly when value changes of quantities correspond to the establishment of media or barriers.

For example, one of the limit points for the height quantity of water in a sink is the value corresponding to the height of the safety drain. This value is a threshold because it is at this height that a medium is established through which a new flow process affects the rate of the water's height. Once threshold values are established in this way, they can be recorded in the value spaces of quantities to support future limit analysis.

For de Kleer and Brown's qualitative reasoning system, the laws describing the behavior of components and the behavior along *conduits* which connect components are represented as *confluences*, or qualitative differential equations. All of the causal mechanism descriptions in the causal modelling system's domain theory are (as argued earlier) easily interpreted as conservation laws. These conservation laws are easily expressed as confluences:

$$\partial iq + \partial dq = 0$$

The causal mechanisms called *transformations* correspond to laws about primitive component models in de Kleer and Brown's domain theory. The causal mechanisms called *propagations*, which describe media, correspond to laws about conduits.

de Kleer and Brown's envisioning system uses confluences to determine the possible behaviors of device components and the possible transitional behaviors along conduits. The possible behaviors of the overall device then are inferred in a bottom-up fashion. The causal models produced by my learning system, appropriately recast in terms of confluences, could support envisioning.

In general, as de Kleer points out, envisioning may generate several possible behaviors for a device. This ambiguity may be due to the loss of information associated with qualitative values, or may reflect inherent indeterminism in the device.

Another kind of reasoning which causal models could be targeted for is fault diagnosis. Knowledge about causal mechanisms can support fault diagnosis. As Davis points out [Davis 84], part of any diagnostic capability lies in knowledge of the *causal pathways* which exist in a domain. In particular, it is knowledge of how *spurious* or *inhibited* causal pathways might arise which can be used to troubleshoot a misbehaving device. My representation for causal mechanism has the right vocabulary for diagnosis: misbehavior due to a spurious causal pathway is explainable in terms of an unexpected medium; misbehavior due to an inhibited causal pathway is explainable in terms of an unexpected barrier.

The domain theory of Davis' hardware troubleshooting system contains knowledge of causal mechanisms which can underlie faults in digital circuits (e.g., shorts, bridges). This domain theory – like the domain theory of my causal modelling system – comes in layers, allowing assumptions about "normal" behavior to be successively relaxed. Each level in the domain theory corresponds to an hypothesis about the increasingly unlikely presence or absence of a causal mechanism. Diagnosis can be seen as the refinement or elaboration of a causal model to explain exceptional behavior.

When a causal model produced by my learning system generates a prediction which is not corroborated by observations, diagnosis can proceed as follows: Employing a single-fault assumption, Davis' *constraint suspension* technique can be used on the network of functional dependencies in the causal model to isolate possible locations of unexpectedly active or inactive causal mechanisms. [1] Then the more detailed levels of the domain theory can be used to generate hypotheses about what exceptional media or barriers might be in place that can explain the misbehavior.

### 8.1.2 Empirical and Analytical Learning

The inductive component of my causal modelling system is similar to other learning systems which learn functional dependencies empirically, of which BACON is best known example [Langley et al 83]. All of these purely inductive systems suffer from two shortcomings: they have no way of determining the relevant features (for my system, the possible causes) and they have no way to justify their hypotheses.

Providing a learning system with a domain theory addresses both of these shortcomings. A domain theory is an implicit statement of what is relevant, and hypotheses derived from a domain theory are justified owing to the independent justification of the theory itself.

The analytical component of my learning system, which makes use of a domain theory, is an example of an *explanation-based learning* system [Mitchell 83, DeJong 83, Mahadevan 85], capable of making relevant, justified generalizations. My system, which also employs schemata, mostly closely resembles that of DeJong.

With the advent of knowledge-intensive, analytical approaches to learning, a research direction which suggests itself is the exploration of ways to combine empirical and analytical learning methods. In my causal modelling system, empirical conjectures can be justified by instantiating a causal mechanism description which explains the empirically noted dependence. Put another way, empirical conjectures can drive an analytical method, providing it with something to explain.

### 8.1.3 Levels Of Abstraction

Even analytically generated causal explanations may prove to be inconsistent, but they can be refined by instantiating a causal mechansism description at a lower level of detail. The more detailed explanation may cover a case which is excluded from the approximate explanation at the higher level.

Reid Smith et al [Smith et al 85] show how the empirically motivated expansion of an explanation (possibly generated by an analytical method) can be focussed by tagging assertions in explanations as *definitional*, *statistical*, *default*, or *approximate*. Definitional assertions are irrefutable axioms;

[1] The technique cannot be expected to work as well with qualitative values as with quantitative values (as in Davis' system), i.e., fewer candidates are eliminated.

the other types of assertions are various kinds of abstractions, any of which can be a source of inconsistency. In my domain theory, all abstractions are of the *approximation* type.

Patil [Patil 81] has addressed the issue of how to shift the level of causal description - either elaborating or abstracting - while maintaining consistency across levels.

### 8.1.4 Learning from Experiments

My learning system has the ability to design experiments both to distinguish competing hypotheses, even partially justified hypotheses, and to further justify a particular hypothesis. Rajamoney et al [Rajamoney et al 85] also discuss the use of experimenting in learning. They outline an experiment designer which determines situations which predict different outcomes for competing hypotheses. As in my learning system, situations are determined by consulting a domain theory which describes relevant dependencies.

## 8.2 The Issues Revisited

Several issues related to the causal modelling competence are addressed in this research. In this section, I summarize proposed solutions and ideas toward solutions for these issues.

- Levels of Abstraction

  Perceptions and causal explanations come at different levels of resolution. The best explanations are those that are consistent with observations, and have some degree of justification. Given a domain theory containing layered approximations, an inconsistent explanation often can be repaired by expanding to a less approximate, and more detailed explanation. The more detailed explanation is more justified because it can handle more exceptional cases. In order to instantiate any particular explanation, it may be necessary to procure perceptions at a finer level of granularity.

- Learning from Experiments

  The ability to design experiments changes a learning system from passive to active. Given that the hypotheses generated by a learning system using an imperfect domain theory always have some empirical basis, the ability to deliberately choose the next example to gather more justification for fewer hypotheses is important. The best experiments are those which are fully instantiable and whose results have unambiguous interpretations. I have shown how experiments can be designed using a constraint back-propagation technique.

- The Domain Theory

  The performance of the causal modelling system depends in part on the correctness of the domain theory. This is a knowledge engineering issue. Three broad classes of causal mechanisms appear in the domain theory, all of which reflect well-known conservation laws from the field of physics. In propagations, stuff is transferred along some medium from one location to another. In transformations, stuff changes form within an object. In field interactions, the motion of an object is relative to its position within a field. The domain theory is neither consistent (because of approximations which abstract away from unusual cases), nor complete.

- Learning Compositions of Causal Mechanisms

  Compositions of causal mechanisms in the domain theory are learned by a standard explanation-based learning technique. Intersections between mechanisms are noted as causal models are constructed and these composed mechanisms are generalized by generalizing their consituent mechanisms while preserving the constraints at the intersections.

- Combining Empirical and Analytical Learning Methods

  The empirical and analytical components of my learning system interact in these interesting ways: Conjectures with a purely or highly empirical basis can be made more justified by employing the analytical method of instantiating causal mechanism descriptions at a more detailed level of explanation. Competing hypotheses, even justified hypotheses, can be distinguished empirically by designing an experiment for which the hypotheses generate incompatible predictions.

- Representing Causality

  Two representations of causality are used by my learning system. One representation is *associationist*, and casts causal relations as functional dependencies between quantities; it is adequate for an inductive approach to causal modelling. The other representation is *mechanistic*, using terms such as medium and barrier; it supports the generation of justified causal explanations which get at why the effect should necessarily follow from the cause. The design of both of these representations was inspired by models of causality from the field of philosophy.

# References


[Davis 84] Davis, Randall, "Diagnostic Reasoning Based on Structure and Behavior," *Artificial Intelligence* **24**, 1984.

[DeJong 83] DeJong, Gerald, "Acquiring Schemata Through Understanding and Generalizing Plans," *8th IJCAI*, Karlsruhe, 462-464, 1983.

[de Kleer & Brown 84] de Kleer, Johan and Brown, John S., "A Qualitative Physics Based on Confluences," *Artificial Intelligence* **24**, 1984.

[Doyle 79] Doyle, Jon, "A Truth Maintenance System," *Artificial Intelligence* **12**, 1979.

[Doyle 84] Doyle, Richard J., "Hypothesizing and Refining Causal Models," MIT Artificial Intelligence Lab AIM-811, 1984.

[Forbus 84] Forbus, Kenneth D., "Qualitative Process Theory," *Artificial Intelligence* **24**, 1984.

[Hayes 79] Hayes, Patrick J., "The Naive Physics Manifesto," in *Expert Systems in the Micro-Electronic Age*, ed. Donald Michie, Edinburgh University Press, Edinburgh, 1979.

[Langley et al 83] Langley, Pat, Gary L. Bradshaw and Herbert A. Simon, "Rediscovering Chemistry with the BACON System," in *Machine Learning – An Artificial Intelligence Approach*, eds. Ryszard S. Michalski, Jaime G. Carbonell and Tom M. Mitchell, Tioga, 1983.

[Mackie 67] Mackie, John L., "Mill's Methods of Induction," in *Encyclopedia of Philosophy*, ed. P. Edwards, New York, 1967.

[Mackie 74] Mackie, John L., *The Cement of the Universe: A Study of Causation*, Oxford University Press, Oxford, 1974.

[Mahadevan 85] Mahadevan, Sridhar, "Verification-Based Learning: A Generalization Strategy for Problem-Reduction Methods," *9th IJCAI*, Los Angeles, 616-623, 1985.

[Michalski 83] Michalski, Ryszard, S., "A Theory and Methodology of Inductive Learning," *Artificial Intelligence* **20**, 1983.

[Mitchell 83] Mitchell, Tom M., "Learning and Problem Solving," Computers and Thought Lecture, *8th IJCAI*, Karlsruhe, 1139-1151, 1983.

[Patil 81] Patil, Ramesh S., "Causal Representation of Patient Illness for Electrolyte and Acid-Base Diagnosis," MIT Laboratory for Computer Science TR-267, 1981.

[Rajamoney et al 85] Rajamoney, Shankar, Gerald DeJong and Boi Faltings, "Towards a Model of Conceptual Knowledge Acquisition Through Directed Experimentation," *9th IJCAI*, Los Angeles, 688-690, 1985.

[Smith et al 85] Smith, Reid G., Howard A. Winston, Tom M. Mitchell and Bruce G. Buchanan, "Representation and Use of Explicit Justifications for Knowledge Base Refinement," *9th IJCAI*, Los Angeles, 673-680, 1985.

[Utgoff 85] Utgoff, Paul E., "Shift of Bias for Inductive Concept Learning," in *Machine Learning - An Artificial Intelligence Approach,* vol. 2, eds. Ryszard S. Michalski, Jaime G. Carbonell and Tom M. Mitchell, Morgan Kaufmann, 1985.

[Weld 84] Weld, Daniel S., "Switching from Discrete to Continuous Process Models to Predict Genetic Activity," MIT Artificial Intelligence Lab TR-793, 1984.

[Winston et al 83] Winston, Patrick H., Thomas O. Binford, Boris Katz and Michael Lowry, "Learning Physical Descriptions from Functional Definitions, Examples, and Precedents," *AAAI-83*, Washington, 433-439, 1983.

# Appendix A

# The Qualitative Calculi

This appendix contains the qualitative calculi which show how the signs and transitions of quantities combine under negation, multiplication, and addition.

## A.1 Calculi for the Signs of Quantities

This section contains the qualitative calculi for the signs of quantities. The three possible signs are 0, +, and −.

### A.1.1 Signs under Negation

| $sign$ | $-sign$ |
|---|---|
| 0 | 0 |
| + | − |
| − | + |

### A.1.2 Signs under Multiplication

| $sign1$ | $sign2$ | $sign1 \times sign2$ |
|---|---|---|
| 0 | 0 | 0 |
| + | 0 | 0 |
| − | 0 | 0 |
| 0 | + | 0 |
| + | + | + |
| − | + | − |
| 0 | − | 0 |
| + | − | − |
| − | − | + |

### A.1.3 Signs under Addition

Note the indeterminism.

| $sign1$ | $sign2$ | $sign1 + sign2$ |
|---|---|---|
| 0 | 0 | 0 |
| + | 0 | + |
| − | 0 | − |
| 0 | + | + |
| + | + | + |
| − | + | 0 + − |
| 0 | − | − |
| + | − | 0 + − |
| − | − | − |

## A.2 Calculi for the Transitions of Quantities

This section contains the qualitative calculi for the transitions of quantities. The thirteen possible transitions are defined below.

| Transition | Old sign | New sign | Direction of change | Abbreviation |
|---|---|---|---|---|
| STAY0 | 0 | 0 | 0 | ST0 |
| STAY+ | + | + | 0 | ST+ |
| STAY- | − | − | 0 | ST- |
| ADD+ | + | + | + | AD+ |
| ADD- | − | − | − | AD- |
| SUBTRACT+ | + | + | − | SB+ |
| SUBTRACT- | − | − | + | SB- |
| ENABLE+ | 0 | + | + | EN+ |
| ENABLE- | 0 | − | − | EN- |
| DISABLE+ | + | 0 | − | DB+ |
| DISABLE- | − | 0 | + | DB- |
| REVERSE+ | + | − | − | RV+ |
| REVERSE- | − | + | + | RV- |

### A.2.1 Transitions under Negation

| *Transition* | $-$*Transition* |
|---|---|
| ST0 | ST0 |
| ST+ | ST- |
| ST- | ST+ |
| AD+ | AD- |
| AD- | AD+ |
| SB+ | SB- |
| SB- | SB+ |
| EN+ | EN- |
| EN- | EN+ |
| DB+ | DB- |
| DB- | DB+ |
| RV+ | RV- |
| RV- | RV+ |

### A.2.2 Transitions under Multiplication

Redundant combinations due to the commutativity of multiplication are not listed.

| *Transition*1 | *Transition*2 | *Transition*1 $\times$ *Transition*2 |
|---|---|---|
| ST0 | ST0 | ST0 |
| ST0 | ST+ | ST0 |
| ST0 | ST- | ST0 |
| ST0 | AD+ | ST0 |
| ST0 | AD- | ST0 |
| ST0 | SB+ | ST0 |
| ST0 | SB- | ST0 |
| ST0 | EN+ | ST0 |
| ST0 | EN- | ST0 |
| ST0 | DB+ | ST0 |
| ST0 | DB- | ST0 |
| ST0 | RV+ | ST0 |
| ST0 | RV- | ST0 |

.
.
.

| *Transition1* | *Transition2* | *Transition1* × *Transition2* |
|---|---|---|
| ST+ | ST+ | ST+ |
| ST+ | ST- | ST- |
| ST+ | AD+ | AD+ |
| ST+ | AD- | AD- |
| ST+ | SB+ | SB+ |
| ST+ | SB- | SB- |
| ST+ | EN+ | EN+ |
| ST+ | EN- | EN- |
| ST+ | DB+ | DB+ |
| ST+ | DB- | DB- |
| ST+ | RV+ | RV+ |
| ST+ | RV- | RV- |
| ST- | ST- | ST+ |
| ST- | AD+ | AD- |
| ST- | AD- | AD+ |
| ST- | SB+ | SB- |
| ST- | SB- | SB+ |
| ST- | EN+ | EN- |
| ST- | EN- | EN+ |
| ST- | DB+ | DB- |
| ST- | DB- | DB+ |
| ST- | RV+ | RV- |
| ST- | RV- | RV+ |
| AD+ | AD+ | AD+ |
| AD+ | AD- | AD- |
| AD+ | SB+ | ST+ AD+ SB+ |
| AD+ | SB- | ST- AD- SB- |
| AD+ | EN+ | EN+ |
| AD+ | EN- | EN- |
| AD+ | DB+ | DB+ |
| AD+ | DB- | DB- |
| AD+ | RV+ | RV+ |
| AD+ | RV- | RV- |
| AD- | AD- | AD+ |
| AD- | SB+ | ST- AD- SB- |
| AD- | SB- | ST+ AD+ SB+ |
| AD- | EN+ | EN- |
| AD- | EN- | EN+ |
| AD- | DB+ | DB- |
| AD- | DB- | DB+ |
| AD- | RV+ | RV- |
| AD- | RV- | RV+ |

| *Transition1* | *Transition2* | *Transition1* × *Transition2* |
|---|---|---|
| SB+ | SB+ | SB+ |
| SB+ | SB- | SB- |
| SB+ | EN+ | EN+ |
| SB+ | EN- | EN- |
| SB+ | DB+ | DB+ |
| SB+ | DB- | DB- |
| SB+ | RV+ | RV+ |
| SB+ | RV- | RV- |
| SB- | SB- | SB+ |
| SB- | EN+ | EN- |
| SB- | EN- | EN+ |
| SB- | DB+ | DB- |
| SB- | DB- | DB+ |
| SB- | RV+ | RV- |
| SB- | RV- | RV+ |
| EN+ | EN+ | EN+ |
| EN+ | EN- | EN- |
| EN+ | DB+ | ST0 |
| EN+ | DB- | ST0 |
| EN+ | RV+ | EN- |
| EN+ | RV- | EN+ |
| EN- | EN- | EN+ |
| EN- | DB+ | ST0 |
| EN- | DB- | ST0 |
| EN- | RV+ | EN+ |
| EN- | RV- | EN- |
| DB+ | DB+ | DB+ |
| DB+ | DB- | DB- |
| DB+ | RV+ | DB+ |
| DB+ | RV- | DB- |
| DB- | DB- | DB+ |
| DB- | RV+ | DB- |
| DB- | RV- | DB+ |
| RV+ | RV+ | ST+ AD+ SB+ |
| RV+ | RV- | ST- AD- SB- |
| RV- | RV- | ST+ AD+ SB+ |

### A.2.3 Transitions under Addition

Redundant combinations due to commutativity are not listed. Note again that additive combinations are much more indeterminite than mutliplicative combinations.

| *Transition1* | *Transition2* | *Transition1* + *Transition2* |
|---|---|---|
| ST0 | ST0 | ST0 |
| ST0 | ST+ | ST+ |
| ST0 | ST- | ST- |
| ST0 | AD+ | AD+ |
| ST0 | AD- | AD- |
| ST0 | SB+ | SB+ |
| ST0 | SB- | SB- |
| ST0 | EN+ | EN+ |
| ST0 | EN- | EN- |
| ST0 | DB+ | DB+ |
| ST0 | DB- | DB- |
| ST0 | RV+ | RV+ |
| ST0 | RV- | RV- |
| ST+ | ST+ | ST+ |
| ST+ | ST- | ST0 ST+ ST- |
| ST+ | AD+ | AD+ |
| ST+ | AD- | AD- SB+ EN- DB+ RV+ |
| ST+ | SB+ | SB+ |
| ST+ | SB- | AD+ SB- EN+ DB- RV- |
| ST+ | EN+ | AD+ |
| ST+ | EN- | SB+ DB+ RV+ |
| ST+ | DB+ | SB+ |
| ST+ | DB- | AD+ EN+ RV- |
| ST+ | RV+ | SB+ DB+ RV+ |
| ST+ | RV- | AD+ EN+ RV- |
| ST- | ST- | ST- |
| ST- | AD+ | AD+ SB- EN+ DB- RV- |
| ST- | AD- | AD- |
| ST- | SB+ | AD- SB+ EN- DB+ RV+ |
| ST- | SB- | SB- |
| ST- | EN+ | SB- DB- RV- |
| ST- | EN- | AD- |
| ST- | DB+ | AD- EN- RV+ |
| ST- | DB- | SB- |
| ST- | RV+ | SB- DB- RV- |
| ST- | RV- | AD- EN- RV+ |

.
.
.

| *Transition1* | *Transition2* | *Transition1* + *Transition2* |
|---|---|---|
| AD+ | AD+ | AD+ |
| AD+ | AD- | ST0 ST+ ST- AD+ AD- SB+ SB- EN+ EN- DB+ DB- RV+ RV- |
| AD+ | SB+ | ST+ AD+ SB+ |
| AD+ | SB- | AD+ SB- EN+ DB- RV- |
| AD+ | EN+ | AD+ |
| AD+ | EN- | SB+ DB+ RV+ |
| AD+ | DB+ | ST+ AD+ SB+ |
| AD+ | DB- | AD+ EN+ RV- |
| AD+ | RV+ | ST+ AD+ SB+ DB+ RV+ |
| AD+ | RV- | AD+ EN+ RV- |
| AD- | AD- | AD- |
| AD- | SB+ | AD- SB+ EN- DB+ RV+ |
| AD- | SB- | ST- AD- SB- |
| AD- | EN+ | SB- DB- RV- |
| AD- | EN- | AD- |
| AD- | DB+ | AD- SB+ EN- DB+ RV+ |
| AD- | DB- | ST- AD- SB- |
| AD- | RV+ | AD- EN- RV+ |
| AD- | RV- | ST- AD- SB- DB- RV- |
| SB+ | SB+ | SB+ |
| SB+ | SB- | ST0 ST+ ST- AD+ AD- SB+ SB- EN+ EN- DB+ DB- RV+ RV- |
| SB+ | EN+ | ST+ AD+ SB+ |
| SB+ | EN- | SB+ DB+ RV+ |
| SB+ | DB+ | SB+ |
| SB+ | DB- | ST+ AD+ SB+ EN+ RV- |
| SB+ | RV+ | SB+ DB+ RV+ |
| SB+ | RV- | ST+ AD+ SB+ EN+ RV- |
| SB- | SB- | SB- |
| SB- | EN+ | SB- DB- RV- |
| SB- | EN- | ST- AD- SB- |
| SB- | DB+ | ST- AD- SB- EN- RV+ |
| SB- | DB- | SB- |
| SB- | RV+ | ST- AD- SB- EN- RV+ |
| SB- | RV- | SB- DB- RV- |
| EN+ | EN+ | EN+ |
| EN+ | EN- | ST0 EN+ EN- |
| EN+ | DB+ | ST+ AD+ SB+ |
| EN+ | DB- | RV- |
| EN+ | RV+ | ST+ AD+ SB+ DB+ RV+ |
| EN+ | RV- | RV- |

| *Transition1* | *Transition2* | *Transition1 + Transition2* |
|---|---|---|
| EN- | EN- | EN- |
| EN- | DB+ | RV+ |
| EN- | DB- | ST- AD- SB- |
| EN- | RV+ | RV+ |
| EN- | RV- | ST- AD- SB- DB- RV- |
| DB+ | DB+ | DB+ |
| DB+ | DB- | STO DB+ DB- |
| DB+ | RV+ | RV+ |
| DB+ | RV- | ST+ AD+ SB+ EN+ RV- |
| DB- | DB- | DB- |
| DB- | RV+ | ST- AD- SB- EN- RV+ |
| DB- | RV- | RV- |
| RV+ | RV+ | RV+ |
| RV+ | RV- | STO ST+ ST- AD+ AD- SB+ SB- EN+ EN- DB+ DB- RV+ RV- |
| RV- | RV- | RV- |